Exploration and Characterisation of the Energy Landscapes of Biomolecules and Hydrated Ions

by

Lewis Conrad Smeeton

A thesis submitted to the

University of Birmingham

for the degree of

DOCTOR OF PHILOSOPHY

School of Chemistry

College of Engineering and Physical Sciences

University of Birmingham

July 2015

Abstract

This study is primarily concerned with the application of energy landscape theory to problems in molecular science. Due to the diverse nature of the research undertaken, the thesis is divided into two parts.

In the first part, the energy landscapes of the BLN-69 and Gō-69 model proteins are investigated using metric disconnectivity graph analysis, which incorporates both the kinetic and geometric structure of an energy landscape. Results are presented which demonstrate the effectiveness with which metric disconnectivity graphs can elucidate the connection between structures in the energy landscape, such as funnels and barriers, with certain structural motifs of the protein, including chain slips and twists in the turn regions.

The second part of the thesis focuses on the energy landscape exploration of micro-hydrated ions. A combination of global optimisation and discrete path sampling is used to study the energy landscapes of five anionic species; sulfate, thiocyanate, perchlorate, ferricyanide and ferrocyanide. For the hydrated sulfate and ferricyanide ions, rigid-body modelling of the molecular species is capable of replicating the physical chemistry of these systems, with both exhibiting a size-dependent suppression of dangling OH bonds, provisionally consistent with infrared photodissociation spectroscopy of size selected hydrated ions. For all the systems studied, it is demonstrated that ions and water molecules modelled as rigid-bodies interacting via pair-wise empirical potentials are capable of exhibiting a wide range of water structure morphologies.

Acknowledgements

I would like to thank my supervisor, Prof. Roy Johnston, for his insight, support and advice throughout the course of my PhD. I would also like to thank Dr. Mark Oakley, who provided a great deal of scientific, technical and life advice, and Dr. Dwaipayan Chakrabarti for his analysis and scientific enquiry, without which I would have been infinitely less productive.

I must also thank Dr. Jacob Stevenson and Dr. Victor Ruehle at the University of Cambridge for their guidance and technical support regarding the various programming and software challenges I undertook during my studies.

This research was funded by the Engineering and Physical Sciences Research Council (EPSRC) under Programme Grant EP/I001352/1. Computational facilities were provided by the MidPlus consortium under EPSRC grant EP/K000128/1 and the University of Birmingham's BlueBEAR HPC service.

I would like to thank the various iterations of the Johnston and Chakrabarti groups for providing a pleasant and enjoyable working environment. I would particularly like to thank Jack Davis, Daniel Morphew, John Hey, Dr. Paul Jennings and Dr. Sridhar Neelamraju. Without their good humour, companionship and vocal opinions my time in Birmingham would have been considerably less fun.

I must thank my parents, Phil and Hazel, for all the support, guidance and encouragement they have shown me through my many years of study. Finally, I thank Hazel for her constant love, kindness and understanding.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Much of the rich and complex behaviour observed in chemical and molecular science arises from the interactions of the many atoms or molecules present in a given system of interest. This behaviour can often span many times scales, and involve large conformational or structural changes as the system evolves. This poses a problem for the theoretical study of these types of systems, because even if the equations of motion are known (and for problems in chemistry they almost inevitably are) finding an analytical solution is typically not possible. Therefore, one must resort to simulation to investigate chemical phenomena, and energy landscape theory is a powerful tool which allows us to do so.

This chapter is a brief introduction to energy landscape theory as it is used and understood in molecular science. Throughout this thesis, potential energy surface and energy landscape shall be used interchangeably.

1.1 Potential Energy Surfaces and the Born-Oppenheimer Approximation

Energy landscape theory is concerned with studying the potential energy surface of a chemical system in order to determine its structural, thermodynamic and kinetic properties [1]. The potential energy surface describes the potential energy of a system as a function of the relative positions and orientations of its constituent parts. Energy landscape theory

is sufficiently general that any collection of bodies can be considered, providing that their interaction energy can be computed. However, at its most fundamental it must be able to treat individual atoms within a molecule or bulk phase, which raises the question: How can one define a potential energy surface for such a system? At the atomic scale, all system properties are derived or computed using a quantum mechanical treatment, and so we start with the time-independent Schrödinger equation

$$\hat{\mathcal{H}}\psi\left(\mathbf{r}_N, \mathbf{r}_n\right) = E_{tot}\psi\left(\mathbf{r}_N, \mathbf{r}_n\right) \tag{1.1}$$

where $\mathbf{r}_N$ and $\mathbf{r}_n$ are the nuclear and electronic coordinates, respectively, E_{tot} is the total energy, $\psi\left(\mathbf{r}_N, \mathbf{r}_n\right)$ is the corresponding eigenstate or wavefunction and $\hat{\mathcal{H}}$ is the molecular Hamiltonian, which takes the general form

$$\begin{aligned}\hat{\mathcal{H}} =& -\sum_I^N \frac{\hbar^2}{2M_I}\nabla_I^2 - \sum_i^n \frac{\hbar^2}{2m}\nabla_i^2 + \frac{e^2}{4\pi\epsilon_0}\sum_I^N\sum_{J>I}^N \frac{Z_I Z_J}{|\mathbf{r}_{N,I} - \mathbf{r}_{N,J}|} \\ &+ \frac{e^2}{4\pi\epsilon_0}\sum_i^n\sum_{j>i}^n \frac{1}{|\mathbf{r}_{n,i} - \mathbf{r}_{n,j}|} - \frac{e^2}{4\pi\epsilon_0}\sum_I^N\sum_i^n \frac{Z_I}{|\mathbf{r}_{N,I} - \mathbf{r}_{n,i}|}\end{aligned} \tag{1.2}$$

$$= \hat{\mathcal{T}_N} + \hat{\mathcal{T}_n} + \hat{\mathcal{U}_{NN}} + \hat{\mathcal{U}_{nn}} + \hat{\mathcal{U}_{Nn}} \tag{1.3}$$

where Z_I is the atomic number (nuclear charge) of the I^{th} nucleus, $-e$ is the electronic charge and $\hat{\mathcal{T}_N}$, $\hat{\mathcal{T}_n}$, $\hat{\mathcal{U}_{NN}}$, $\hat{\mathcal{U}_{nn}}$ and $\hat{\mathcal{U}_{Nn}}$ are the nuclear and electronic kinetic energy operators and nuclear-nuclear, electron- electron and nuclear-electron interaction potential operators, respectively. In order to define a potential energy surface for such a system, one must employ the Born-Oppenheimer (BO) approximation to separate the nuclear and electronic motion [2, 3, 4]. The BO approximation is central to quantum chemistry, and recognises that the atomic nuclei are considerably more massive than the electrons (the proton to electron mass ratio is approximately 1836:1), and thus the nuclear and electronic wave functions can be adiabatically decoupled. This is achieved using a separation-of-variables approach and assuming a separable ansatz of the form

$$\psi\left(\mathbf{r}_N, \mathbf{r}_n\right) = \phi\left(\mathbf{r}_n; \mathbf{r}_N\right)\chi\left(\mathbf{r}_N\right) \tag{1.4}$$

where $\phi(\mathbf{r}_n;\mathbf{r}_N)$ and $\chi(\mathbf{r}_N)$ are the electronic and nuclear wavefunctions, respectively. The notation $\phi(\mathbf{r}_n;\mathbf{r}_N)$ indicates that the electronic wavefunction is a function of the electronic coordinates, $\mathbf{r}_n$, and that it depends *parametrically* on the nuclear coordinates, $\mathbf{r}_N$. One can then substitute Equation 1.4 into Equation 1.1 to give

$$\begin{aligned}\hat{\mathcal{H}}\phi(\mathbf{r}_n;\mathbf{r}_N)\chi(\mathbf{r}_N) = &-\sum_I^N \frac{\hbar^2}{2M_I}\phi(\mathbf{r}_n;\mathbf{r}_N)\nabla_I^2\chi(\mathbf{r}_N) \\ &-\sum_I^N \frac{\hbar^2}{2M_I}\left[\chi(\mathbf{r}_N)\nabla_I^2\phi(\mathbf{r}_n;\mathbf{r}_N) + 2\nabla_I\chi(\mathbf{r}_N)\nabla_I\phi(\mathbf{r}_n;\mathbf{r}_N)\right] \\ &+\left[\hat{\mathcal{T}}_n + \hat{\mathcal{U}_{NN}} + \hat{\mathcal{U}_{nn}} + \hat{\mathcal{U}_{Nn}}\right]\phi(\mathbf{r}_n;\mathbf{r}_N)\chi(\mathbf{r}_N) \\ = &\, E_{tot}\phi(\mathbf{r}_n;\mathbf{r}_N)\chi(\mathbf{r}_N)\end{aligned} \tag{1.5}$$

where only the nuclear kinetic operator has been expanded. Terms which involve the derivative of the electronic wavefunction with respect to the nuclear coordinates (i.e. $\nabla_I^2\phi(\mathbf{r}_n;\mathbf{r}_N)$ and $\nabla_I\phi(\mathbf{r}_n;\mathbf{r}_N)$) are the only means by which the nuclear dynamics can cause an electronic transition. These non-adiabatic coupling terms can be neglected due to the discrepancy between the electronic and nuclear mass, making Equation 1.5 separable. Equation 1.5 can then be rearranged to give

$$\frac{\left[\hat{\mathcal{T}}_n + \hat{\mathcal{U}_{nn}} + \hat{\mathcal{U}_{Nn}}\right]\phi(\mathbf{r}_n;\mathbf{r}_N)}{\phi(\mathbf{r}_n;\mathbf{r}_N)} = E_{tot} - \frac{\left[\hat{\mathcal{T}}_N + \hat{\mathcal{U}_{NN}}\right]\chi(\mathbf{r}_N)}{\chi(\mathbf{r}_N)} \tag{1.6}$$

where the electronic coordinates appear on the left hand side of Equation 1.6 only. The left hand side can therefore be equated to a function of the nuclear coordinates only and expressed as an eigenvalue problem

$$\left[\hat{\mathcal{T}}_n + \hat{\mathcal{U}_{nn}} + \hat{\mathcal{U}_{Nn}}\right]\phi(\mathbf{r}_n;\mathbf{r}_N) = \hat{\mathcal{H}}_e\phi(\mathbf{r}_n;\mathbf{r}_N) = E_e(\mathbf{r}_N)\phi(\mathbf{r}_n;\mathbf{r}_N) \tag{1.7}$$

where $\hat{\mathcal{H}}_e$ is the electronic Hamiltonian. The electronic eigenvalue problem above has a spectrum of solutions and corresponding eigenvalues, $E_e(\mathbf{r}_N)$, which form a set of electronic, or Born-Oppenheimer, surfaces. It is upon these surfaces that the nuclear eigenvalue problem

is solved

$$\left[\hat{\mathcal{T}}_N + \hat{\mathcal{U}_{NN}} + E_e\left(\mathbf{r}_N\right)\right]\chi\left(\mathbf{r}_N\right) = E_{tot}\chi\left(\mathbf{r}_N\right) \tag{1.8}$$

the solutions of which yield the rotational and vibrational motions of the nuclei and their associated energies for a given electronic surface. The total atomic interaction potential is a combination of an electronic surface and the nuclei-nuclei interaction

$$\left[\hat{\mathcal{U}_{NN}} + E_e\left(\mathbf{r}_N\right)\right]\chi\left(\mathbf{r}_N\right) = U\left(\mathbf{r}_N\right)\chi\left(\mathbf{r}_N\right) \tag{1.9}$$

where $U(\mathbf{r}_N)$ is the *potential energy surface* for a given electronic surface. In this thesis we shall only consider systems in the electronic ground state, and from here on we shall drop the subscript, so that $U(\mathbf{r}_N) \equiv U(\mathbf{r})$.

The BO approximation provides a rigorous foundation upon which the potential energy surface can be defined. The derivation presented above implies that in order to compute $U(\mathbf{r})$ one must solve Equation 1.9. In practice such a calculation would be undesirable due to the time and resources it would require, and so cheaper, more approximate methods are typically employed. Throughout this thesis we compute potentials using both quantum chemical calculations at the density functional theory (DFT) level and empirical potentials, depending on the required accuracy, available resources, desired sampling and physical relevance.

With a potential energy surface defined, one can begin to study its key features, which are discussed below in Section 1.2. As an aside, utilising the potential energy surface is not limited to energy landscape theory; it is central to most simulation techniques which investigate chemical and molecular problems, including molecular dynamics [5, 6] and Monte Carlo simulation [7].

1.2 Stationary Points on the Potential Energy Surface

The key features of a potential energy surface are the set of stationary points on the surface $\{\mathbf{r}_\alpha\}$ which satisfy

$$\nabla U\left(\mathbf{r}_i\right) = \mathbf{0}, \qquad \mathbf{r}_i \in \{\mathbf{r}_\alpha\} \tag{1.10}$$

where $\nabla = \left(\frac{\partial}{\partial x_1}, \ldots, \frac{\partial}{\partial x_N}\right)$ in $\mathbb{R}^N$ and $\mathbf{0}$ is the null vector. It is the stationary points on the surface which energy landscape theory is primarily concerned with locating and sampling. The importance of stationary points on a potential landscape can be understood by considering the relationship between potential and force. For a landscape $U(\mathbf{r})$, the force acting on the point $\mathbf{r}$, $\mathbf{F}(\mathbf{r})$, is proportional to the derivative of $U(\mathbf{r})$ with respect to $\mathbf{r}$

$$\mathbf{F}(\mathbf{r}) = -\nabla U(\mathbf{r}) \tag{1.11}$$

so points which are locally flat (i.e. stationary points) are points at which the total force sums to zero. Thus the set of stationary points $\{\mathbf{r}_\alpha\}$ on $U(\mathbf{r})$ defined as the solutions to Equation 1.10 are also all the points on the surface which are at mechanical equilibrium. However, not all stationary points are equal, and not all stationary points are of equal importance when attempting to understand the physical and chemical properties of a system. Although Equation 1.10 guarantees that a stationary point will be in mechanical equilibrium, it offers no guarantee that a stationary point will also be *mechanically stable*.

Mechanical stability describes the manner in which a system evolves through time according to its equations of motion when in the vicinity of a stationary point. As an illustration of the principle, one can consider a system initially prepared in some stationary state which is then subject to a small perturbation. If the initial state is mechanically stable, then the governing equations of motion will act to return the system to its initial state, and the system will oscillate indefinitely about the stationary point. However, if the stationary point is unstable, then the equations of motion will not return the system to its initial state, and the two will instead diverge. For this reason, it is the mechanically stable stationary points which one wishes to sample when trying to compute thermodynamic properties, as it is in the vicinity of these points which the system will spend the most time. This is particularly true for low temperature problems, where the dynamical behaviour is well described as a system vibrating about a low energy stationary point.

So how does one determine the local mechanical stability of a stationary point? One can gain insight by first determining the local curvature of the stationary point, and then studying the

forces acting on the system about the stationary point. A useful object for understanding the shape and curvature of a surface $U(\mathbf{r})$ is the Hessian, a matrix of second partial derivatives of U with respect to $\mathbf{r}$

$$H_{ij}\left(\mathbf{r}\right) = \frac{\partial^2 U\left(\mathbf{r}\right)}{\partial r_i \partial r_j} \tag{1.12}$$

The eigenvectors and eigenvalues of the Hessian have a special significance when the Hessian is evaluated at a stationary point

$$\mathbf{H}\left(\mathbf{r}_0\right)\mathbf{b}_\alpha = \varepsilon_\alpha^2 \mathbf{b}_\alpha \tag{1.13}$$

with the eigenvectors, $\mathbf{b}_\alpha$, and eigenvalues, ε_α^2 of $\mathbf{H}\left(\mathbf{r}_0\right)$ being related to the *normal mode* vibrations and the square of the vibrational frequencies, respectively, of the stationary point. As an aside, to compute the normal modes and normal mode frequencies, one must instead use the *mass-weighted* Hessian, $\mathbf{h}$, such that $h_{\alpha\beta} = H_{\alpha\beta}/\sqrt{M_\alpha M_\beta}$. The shape of the potential landscape in the vicinity of a stationary point can be inspected by expanding $U(\mathbf{r})$ in a Taylor series

$$U\left(\mathbf{r}\right) = U\left(\mathbf{r}_0\right) + \nabla U\left(\mathbf{r}_0\right)^T\left(\mathbf{r}-\mathbf{r}_0\right) + \frac{1}{2}\left(\mathbf{r}-\mathbf{r}_0\right)^T \mathbf{H}\left(\mathbf{r}_0\right)\left(\mathbf{r}-\mathbf{r}_0\right) + \ldots \tag{1.14}$$

where $\mathbf{H}\left(\mathbf{r}_0\right) = \mathbf{H}\left(\mathbf{r}\right)\big|_{\mathbf{r}_0}$ is the Hessian of $U\left(\mathbf{r}\right)$ at the point $\mathbf{r} = \mathbf{r}_0$. Setting $\mathbf{r}_0$ to be a stationary point and restricting $\mathbf{r}$ to the immediate vicinity of $\mathbf{r}_0$ allows one to neglect the first derivative term in Equation 1.14 and to truncate the Taylor series at second order

$$U\left(\mathbf{r}\right) = U\left(\mathbf{r}_0\right) + \frac{1}{2}\left(\mathbf{r}-\mathbf{r}_0\right)^T \mathbf{H}\left(\mathbf{r}_0\right)\left(\mathbf{r}-\mathbf{r}_0\right) \tag{1.15}$$

Substituting Equation 1.15 into Equation 1.11 and evaluating the derivative on the right hand side gives

$$\mathbf{F}(\mathbf{r}) = -\mathbf{H}\left(\mathbf{r}_0\right)\left(\mathbf{r}-\mathbf{r}_0\right) \tag{1.16}$$

which can be simplified further by expanding $(\mathbf{r} - \mathbf{r}_0)$ as a sum of Hessian basis vectors

$$(\mathbf{r} - \mathbf{r}_0) = \sum_{\alpha} b_\alpha \mathbf{b}_\alpha \tag{1.17}$$

and substituting Equation 1.17 into Equation 1.16

$$\mathbf{F}(\mathbf{r}) = -\mathbf{H}(\mathbf{r}_0) \sum_{\alpha} b_\alpha \mathbf{b}_\alpha = -\sum_{\alpha} \varepsilon_\alpha^2 b_\alpha \mathbf{b}_\alpha \tag{1.18}$$

Equation 1.18 now describes the force acting on a system when near to a stationary point. If $\varepsilon_\alpha^2 > 0$ for all α, then the force points back towards the stationary point, and so acts as a restoring force. Therefore, the stationary point which as all positive Hessian eigenvalues is mechanically stable. If one or more of the eigenvalues is negative (i.e. $\varepsilon_\alpha^2 < 0$) then $\mathbf{F}(\mathbf{r})$ will not be restorative, and will instead accelerate the system away from the stationary point making it unstable. A stationary point can be classified by the number of negative eigenvalues its Hessian has; a stationary point with n negative eigenvalues is a rank n stationary point.

Stationary points with no negative Hessian eigenvalues (rank 0) are *minima*; any slight geometric perturbation will increase the energy of the system, and induce a restorative force. A good understanding of the distribution of minima across a surface is vital for determining system properties, especially if one is interested in studying the low temperature regime (where dynamical effects will be less pronounced). Due to their importance, a great deal of time has been spent on developing numerical methods which will quickly and efficiently locate minima on the potential surface, some of which will be discussed in Chapter 2.

However, it is not just minima which are important for predicting system behaviour. If one wishes to determine *kinetic* properties (for example, the rate and mechanism by which the system moves from one minima to another), then one must locate the *transition states* that lie between them. The Murrell-Laidler theorem implies that the lowest energy path between two minima must, at its highest point, pass through a saddle point of rank 1 [8]. Therefore, if one wishes to determine the transition pathway between two states, one must sample rank 1 saddle points, or transition states, as well as the minima themselves. Throughout

Minimum Transition State

Figure 1.1: Two-dimensional representation of a rank 0 stationary point (minimum, left) and rank 1 stationary point (transition state, right).

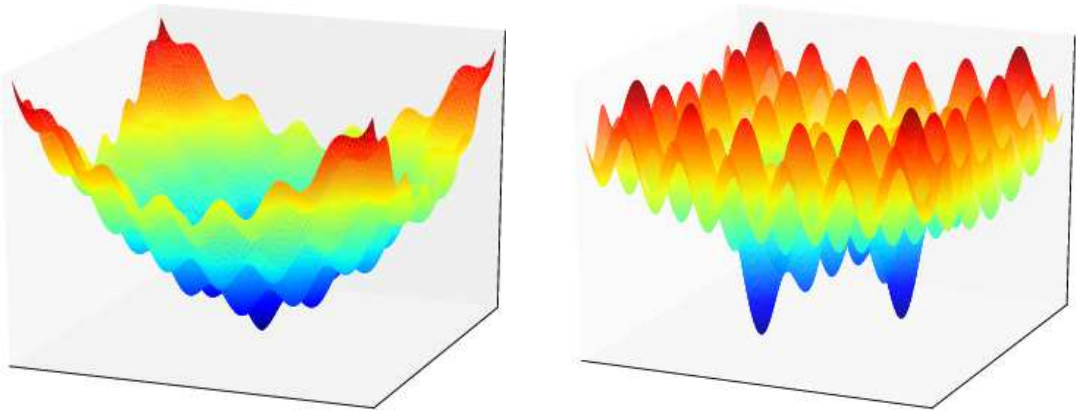
this thesis the terms transition state and rank 1 saddle point shall be used interchangeably. Higher order saddle-points do exist, and methods exist to sample them [9]. However, with the exception of some edge cases (such as molecules with linear transition states which poses more than one negative Hessian eigenvalue [10]), it is not clear if these higher rank stationary points play any great factor in determining the chemical or physical behaviour of a system.

1.3 Topography of the Potential Energy Surface

The topography of a potential landscape describes the distribution of minima and transition states across the surface, and is a significant factor in determining a system's kinetic and thermodynamics properties. By studying and sampling these stationary points one can gain appreciable insight ones system of interest.

When discussing landscape topographies, it is often convenient to think in terms of basins and barriers. Basins, or basins of attraction, are the regions of space about a minimum which are connected to it via a steepest descent minimisation. Barriers, or kinetic barriers, typically refer to the energy of the transition state which joins (or separates) two basins. Super-basins contain a number of individual basins which are kinetically accessible to one another via low barriers.

Broadly speaking, landscape topographies can be divided into two categories; funnelled landscapes and frustrated landscapes (see Figure 1.2 for illustrative examples). Funnelled landscapes are characterised by a single low energy minimum connected to higher energy minima on the surface by asymmetric kinetic barriers [11, 12, 13]. A barrier is considered



A Funnelled Landscape A Frustrated Landscape

Figure 1.2: Two-dimensional representation of a funnelled landscape (left) and a frustrated landscape (right). The frustration in the landscape is due to the two low energy minima of comparable energy separated by a large barrier.

asymmetric if a minimum on side of the barrier is lower in energy than the other (a symmetric barrier would connect two isoenergetic minima). With the exception of the lowest energy minimum, the basins in a funnel are shallow, meaning that the system cannot become kinetically arrested in a metastable state. In effect, the funnel forms a single super-basin. Funnelled landscapes are associated with structure-seeking systems such as proteins or crystals. For these systems, a well defined conformation or structure (which corresponds to the global minimum at the bottom of the funnel) exists, which is readily accessible to the system and can be reached in a reasonable time scale.

The frustrated landscape covers a much greater range of kinetic behaviour, though they all share one characteristic in common; the presence of metastable states on the landscape. At the least frustrated end of the spectrum, this could include a single deep basin on an otherwise funnelled landscape which acts to transiently trap a system whilst on the way to the global minimum. A more frustrated landscape could include a number of low energy minima energetically competitive with the global minimum, each occupying their own funnel. Such an arrangement could lead to occupation probability spread equally between the different metastable states, with slow interconversion rates between the funnels. The most extreme case are glassy systems (so called because they are associated with glass forming materials). Glassy systems exhibit extremely frustrated landscapes, absent of any kinetic funnels and with a large number of energetically degenerate minima separated by large barriers [14, 15].

Chapter 2

Methodology

Details of methodology common to all parts of the thesis.

2.1 Locating Stationary Points on the Potential Energy Surface

As discussed in Section 1.2, the stationary points on the potential energy surface, $U(\mathbf{r})$, are defined as solutions to Equation 1.10, which for all but the simplest cases must be solved numerically. The literature is replete with methods, schemes and procedures for optimising molecular geometries to either minima or transition states [16, 17, 18]. The following section does not contain an exhaustive review of optimising techniques, instead giving a brief overview of the methods used by the GMIN suite [19, 20, 21], pele [22] and NWChem software packages [23] used for computing results in this thesis.

2.1.1 Local Optimisation

Newton-Raphson and Quasi-Newton methods are the most efficient and widely used methods for performing local molecular geometry optimisation. The premise behind these procedures is that the potential energy surface about a minimum can be treated harmonically. To begin,

we re-write Equation 1.14 as

$$U\left(\mathbf{r}_0 + \Delta\mathbf{r}\right) = U\left(\mathbf{r}_0\right) + \mathbf{g}_0^T \Delta\mathbf{r} + \frac{1}{2}\Delta\mathbf{r}^T \mathbf{H}_0 \Delta\mathbf{r} \tag{2.1}$$

where $\mathbf{r}_0$ is some initial point, $\Delta\mathbf{r}$ is a displacement, $\mathbf{g}_0 = \nabla U\left(\mathbf{r}\right)\big|_{\mathbf{r}_0}$ and $\mathbf{H}_0 = \mathbf{H}\left(\mathbf{r}\right)\big|_{\mathbf{r}_0}$. Choosing $\Delta\mathbf{r}$ such that $\mathbf{r}_0 + \Delta\mathbf{r}$ is a minimum and differentiating Equation 2.1 with respect to the coordinates yields

$$\mathbf{0} = \mathbf{g}_0 + \mathbf{H}_0 \Delta\mathbf{r} \tag{2.2}$$

which when rearranged gives an expression for the displacement to the minimum

$$\Delta\mathbf{r} = -\mathbf{H}_0^{-1}\mathbf{g}_0 \tag{2.3}$$

where $\Delta\mathbf{r}$ is called the *Newton step*.

If the anharmonicity at $\mathbf{r}_0$ is non-negligible (for example, if we start a long way from the local minimum, or if the system is very anharmonic) several iterations of Equation 2.3 may be required

$$\mathbf{r}_{i+1} = \mathbf{r}_i - \mathbf{H}_i^{-1}\mathbf{g}_i \tag{2.4}$$

where $\Delta\mathbf{r} = \mathbf{r}_{i+1} - \mathbf{r}_i$ and $i = 1, 2, 3, \ldots$ is the iteration index. In the Newton-Raphson scheme, the Hessian is re-computed at each iteration, which becomes increasingly expensive and inefficient as system size increases. In Quasi-Newton methods this bottle-neck is removed by using an approximate Hessian instead, which is updated at each step in the procedure. Providing that the Newton condition is satisfied

$$\Delta\mathbf{r} = \Delta\mathbf{H}^{-1}\Delta\mathbf{g} \tag{2.5}$$

where $\Delta\mathbf{H}^{-1} = \mathbf{H}_{i+1}^{-1} - \mathbf{H}_i^{-1}$ is the update to the inverse Hessian and $\Delta\mathbf{g} = \mathbf{g}_{i+1} - \mathbf{g}_i$, any updating scheme can be used. A successful and commonly used scheme is the Broyden-

Fletcher-Goldfarb-Shanno (BFGS) update [24, 25, 26, 27]

$$\mathbf{H}_{i+1}^{-1} = \mathbf{H}_i^{-1} + \frac{\Delta\mathbf{r}\Delta\mathbf{r}^T}{\Delta\mathbf{r}^T\Delta\mathbf{g}} - \frac{\mathbf{H}_i^{-1}\Delta\mathbf{g}\Delta\mathbf{g}^T\mathbf{H}_i^{-1}}{\Delta\mathbf{g}^T\mathbf{H}_i^{-1}\Delta\mathbf{g}} \tag{2.6}$$

which is guaranteed to be positive definite (i.e. the update has all positive eigenvalues) ensuring that the optimisation step is always towards a minimum, and not a transition state or other stationary point. The GMIN suite, pele and NWChem each use a limited memory version of BFGS [28] (L-BFGS), which replaces storing the full (inverse) Hessian matrix (an $\mathcal{O}(n^2)$ requirement for n variables) with the $\Delta\mathbf{r}$ and $\Delta\mathbf{g}$ for a limited number of previous steps (an $\mathcal{O}(n)$ requirement). If $U(\mathbf{r})$ is harmonic, Equation 2.3 is exact, otherwise the step-size, $|\Delta\mathbf{r}|$, must be amended somehow. The NWChem package achieves this with a modified CSRCH line search algorithm [29], whilst pele and the GMIN suite accept the default size and direction of $\Delta\mathbf{r}$, but with a constraint that if the energy rises by more than a critical amount, the step-size is reduced until the condition is satisfied.

2.1.2 Transition State Searching

Determining the location of transition states is considerably harder than for local minima. The Hessian matrix evaluated at a transition state has one *negative* eigenvalue, meaning that along the direction of the corresponding eigenvector the transition state is a local *maximum* and is akin to a structure balancing on a knife edge. A two-step process was used to find transition states in this thesis; a doubly nudged elastic band method (DNEB) [30] was used to find candidate transition state structures which were then refined using hybrid eigenvector-following [31, 32].

DNEB is a chain of states method [33, 34], which locates transition states by searching for minimum energy pathways between local minima. In a chain of states calculation, the pathway between two end-points is discretised into a series of points or images, which subject to a constrained optimisation will (hopefully) converge onto a minimum energy pathway. One

can write a simple objective function for such a calculation as

$$S(\mathbf{R}_1, \ldots, \mathbf{R}_{M-1}) = U^{path} + U^{spring} \tag{2.7}$$

$$U^{path} = \sum_{i=0}^{M} U(\mathbf{R}_i) \tag{2.8}$$

$$U^{spring} = \sum_{i=1}^{M} \frac{1}{2} k^{spring} (\mathbf{R}_i - \mathbf{R}_{i-1})^2 \tag{2.9}$$

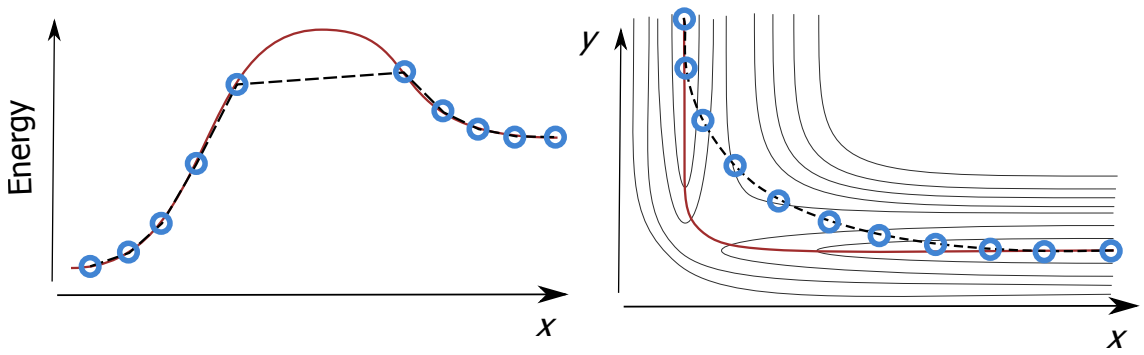
where $\mathbf{R}_i$ is the structure of the i^{th} image along the path, U^{path} is the sum of the energies of the images along the path, U^{spring} is a harmonic constraint which acts to pull images along the reaction path, and k^{spring} is the spring constant between adjacent images. An initial pathway between the end images $\mathbf{R}_0$ and $\mathbf{R}_M$ can be chosen (typically a linear interpolation between the two will suffice [35]), from which the minimum energy pathway can then be estimated by optimising $S(\mathbf{R}_1, \ldots, \mathbf{R}_{M-1})$ whilst keeping the end images $\mathbf{R}_0$ and $\mathbf{R}_M$ fixed. Two comments on the optimisation; the total force acting on the chain has $M \times \eta$ degrees of freedom, where η is the number of atomic degrees of freedom per image, so optimising a chain of images is much more expensive than optimising a single structure. Also, without U^{spring} acting as a constraint, the images would optimise to their nearest local minimum and no transition state candidates would be sampled. This harmonic constraint is the elastic band part of DNEB.

If left unmodified, the above formulation does not perform well. Instead of converging on the minimum energy pathway, the chain of images will corner-cut in regions of high path curvature and slide-down from the transition state regions towards the bottom of basins [35] (see Figure 2.1). We can understand both these phenomena by considering the forces acting on the ith image of our chain

$$\mathbf{F}_i = \mathbf{F}_i^{path} + \mathbf{F}_i^{spring} \tag{2.10}$$

$$\mathbf{F}_i^{path} = -\nabla U(\mathbf{R}_i) \tag{2.11}$$

$$\mathbf{F}_i^{spring} = k^{spring}(\mathbf{R}_{i+1} - \mathbf{R}_i) - k^{spring}(\mathbf{R}_i - \mathbf{R}_{i-1}) \tag{2.12}$$



(a) Images slipping from transition state (b) Corner Cutting in a region of high curvature

Figure 2.1: Schematic of slipping down (a) and corner cutting (b).

and noting that $\mathbf{F}_i^{path}$ and $\mathbf{F}_i^{spring}$ can be resolved into components parallel and perpendicular to the chain

$$\mathbf{F}_i^{path\parallel} = \left(\mathbf{F}_i^{path} \cdot \tilde{\boldsymbol{\tau}}_i\right) \tilde{\boldsymbol{\tau}}_i \tag{2.13}$$

$$\mathbf{F}_i^{path\perp} = \mathbf{F}_i^{path} - \mathbf{F}_i^{path\parallel} \tag{2.14}$$

and similarly for $\mathbf{F}_i^{spring}$, where $\tilde{\boldsymbol{\tau}}_i$ is the unit tangent to the path at the i^{th} image. In regions of high path curvature $\mathbf{F}_i^{spring\perp}$ is significant, preventing the images from closely following the minimum energy path and resulting in corner cutting. When the path crosses a transition state, $\mathbf{F}_i^{path\parallel}$ pulls images away from the top of the barrier, causing the path to densely sample the bottom of basins but only sparsely sample the transition state regions [35]. These problems cannot simply be fixed by altering the value of k^{spring}, as stiff bands will prevent slipping down but aggravate corner cutting, and loose bands will do the opposite. Instead, the remedy is to prevent certain components of $\mathbf{F}_i$ from acting on the chain of images by projecting them out. This process of projecting out is refereed to as nudging, and in DNEB, two rounds of nudging are performed. The first round is equivalent to the (singly) nudged elastic band method (NEB) [35, 36, 37, 38]

$$\mathbf{F}_i^{NEB} = \mathbf{F}_i^{path\perp} + \mathbf{F}_i^{spring\parallel} \tag{2.15}$$

where $\mathbf{F}_i^{NEB}$ is the NEB force acting on the ith image in the chain. In NEB, only the spring force parallel to the chain is used and all else is discarded. The stability and rate

of convergence of the chain to the minimum energy pathway can be further improved by retaining a portion of the spring force perpendicular to the chain. This is done in DNEB by introducing a second nudge, modifying the force further

$$\mathbf{F}_i^{DNEB} = \mathbf{F}_i^{NEB} + \left(\mathbf{F}_i^{spring\perp} - \left(\mathbf{F}_i^{spring} \cdot \hat{\boldsymbol{F}}_i^{path}\right)\hat{\boldsymbol{F}}_i^{path}\right) \tag{2.16}$$

where $\mathbf{F}_i^{DNEB}$ is the DNEB force acting on the ith image in the chain and $\hat{\boldsymbol{F}}_i^{path}$ is the unit vector of the path force. By only using the component of the perpendicular spring force which is also orthogonal to the path force, kinks in the chain are straightened out (which improves convergence) without encouraging corner cutting.

OPTIM and pele both optimise the chain of images in a DNEB calculation using an l-BFGS minimisation routine.

Once a chain has converged on to a minimum energy pathway, transition state candidates can be identified for further refinement using a hybrid eigenvector-following routine. Images are denoted as candidate transition states if they are higher in energy than their two nearest neighbours, so that the ith image is a candidate if $U(\mathbf{R}_i) > U(\mathbf{R}_{i\pm1})$.

Hybrid eigenvector-following is a geometry optimisation routine which locally optimises to a maximum along one degree of freedom (the reaction coordinate of the transition state) whilst simultaneously optimising to a local minimum in the others. The implementation of this routine in pele and OPTIM consists of three stages:

1. The lowest non-zero eigenvalue, ε^2_{min}, and eigenvector, $\mathbf{b}_{min}$, of the Hessian $\mathbf{H}$ are computed.

2. An uphill step parallel to $\mathbf{b}_{min}$ is taken.

3. The coordinates orthogonal to $\mathbf{b}_{min}$ are minimised with respect to $U(\mathbf{r})$.

In the first stage, $\mathbf{b}_{min}$ is treated as an approximation to the reaction coordinate of the transition state undergoing refinement, and is computed by employing a variational approach to minimise the Rayleigh-Ritz ratio [31, 39]

$$\varepsilon^2(\mathbf{r}) = \frac{\mathbf{r}^T\mathbf{H}\mathbf{r}}{\mathbf{r}^2} \tag{2.17}$$

When $\mathbf{r} = \mathbf{b}_{min}$, Equation 2.17 is minimal and takes the value $\varepsilon^2(\mathbf{b}_{min}) = \varepsilon^2_{min}$. (Note. Computing $\mathbf{b}_{min}$ amounts to finding the vector along which the second derivative of $U(\mathbf{r})$ is a minimum). In the interest of efficiency, the second derivatives are approximated as the central difference of $U(\mathbf{r})$

$$\varepsilon^2(\mathbf{r}) \approx \frac{U(\mathbf{r}_0 + \xi\mathbf{r}) - 2U(\mathbf{r}_0) + U(\mathbf{r}_0 - \xi\mathbf{r})}{(\xi\mathbf{r})^2} \tag{2.18}$$

where $\xi \ll 1$ and

$$\frac{\partial\varepsilon^2}{\partial\mathbf{r}} \approx \frac{\nabla U(\mathbf{r}_0 + \xi\mathbf{r}) - \nabla U(\mathbf{r}_0 - \xi\mathbf{r})}{\xi\mathbf{r}^2} - \frac{2\varepsilon^2\mathbf{r}}{\mathbf{r}^2} \tag{2.19}$$

In the second stage, the $\mathbf{b}_{min}$ calculated in Stage 1 is used to direct a step uphill. This stage in the geometry optimisation can be understood by considering the Newton-Raphson step (see Equation 2.3) in the basis of Hessian eigenvectors $\{\mathbf{b}_1, \ldots, \mathbf{b}_{3N}\}$

$$\Delta\mathbf{r} = -\sum_{\alpha}^{3N} \frac{f_\alpha \mathbf{b}_\alpha}{\varepsilon^2_\alpha} \tag{2.20}$$

where f_α is the component of the gradient of U along $\mathbf{b}_\alpha$. Inspection of the energy change associated with the step

$$\Delta U = -\sum_{\alpha}^{3N} \frac{f^2_\alpha}{2\varepsilon^2_\alpha} \tag{2.21}$$

reveals that terms with $\varepsilon^2_\alpha > 0$ act to lower the energy, whilst terms with $\varepsilon^2_\alpha < 0$ do the opposite. Providing that the number of negative eigenvalues did not change from one step to the next, then optimising by Newton's method would be sufficient to converge to the desired transition state. However, this is not guaranteed, and extra care has to be taken to ensure the calculation runs as expected. The eigenvector-following approach [40, 41] solves this by using the method of Lagrange multipliers to constrain the number of negative eigenvalues. By making an appropriate choice for the Lagrange multipliers [42, 43], one arrives at the following step size

$$\Delta\mathbf{r} = \sum_{\alpha}^{3N} \frac{\pm 2 f_\alpha \mathbf{b}_\alpha}{|\varepsilon^2_\alpha| \left(1 + \sqrt{1 + 4f^2_\alpha/\varepsilon^4_\alpha}\right)} \tag{2.22}$$

where the plus and minus are for uphill and downhill steps respectively. In the vicinity of a stationary point, Equation 2.22 becomes the Newton-Raphson step. In practice Equation 2.22 is used to make an uphill step only, and the components of $\Delta\mathbf{r}$ which move downhill are neglected, deferring downhill motion to Stage 3.

In the third stage, $U(\mathbf{r})$ is minimised with respect to the coordinates orthogonal to $\mathbf{b}_{min}$, by projecting out the components of $\nabla U(\mathbf{r})$ parallel to $\mathbf{b}_{min}$.

In Stages 1 and 3 local optimisations of $\varepsilon^2(\mathbf{r})$ and $U(\mathbf{r})$ respectively are performed with respect to atomic coordinates. In OPTIM and pele, both these optimisations are performed using an L-BFGS minimisation routine.

2.2 Global Optimisation

It would be preferable to identify the energy and geometry of every local minimum on the potential energy surface for a given system (i.e. solve Equation1.10 for all $\mathbf{r}_0$ with all-positive Hessian eigenvalues). However, the number of minima on a landscape grows exponentially with the number of dimensions [44, 45, 46], so determining all such structures becomes unfeasible for all but the smallest systems sizes. In such cases, one typically concerns oneself with locating the *lowest* energy minima, and in particular the *global* minimum, as it is these structures which contribute the most to the systems equilibrium thermodynamic properties. To achieve this, one must employ a *global* optimisation strategy to conduct the search. The ease with which the global minimum can be located depends upon the size and nature of the search space in question, but in general is highly non-trivial and requires the use of a sophisticated optimisation strategy. Below we discuss one such method which has been used to collect the data presented in this thesis, Basin-Hopping Monte Carlo.

2.2.1 Basin-Hopping Monte Carlo

Basin-Hopping Monte Carlo is an effective and commonly used global optimisation strategy for problems in molecular science [47, 48, 49]. The premise behind the basin-hopping method is straightforward: if the barriers separating minima on a landscape are removed, then the

rate at which that landscape is explored (and thus the likelihood of locating the global minimum) will increase. Basin-hopping achieves this by introducing the transformation

$$\tilde{U}(\mathbf{r}) = \min\{U(\mathbf{r})\} \tag{2.23}$$

where min maps every point on $U(\mathbf{r})$ to the local minimum of the basin in which it resides, and $\tilde{U}(\mathbf{r})$ is the resultant "staircase" surface. The transformed surface $\tilde{U}(\mathbf{r})$ preserves the energies and coordinates of all minima on $U(\mathbf{r})$ whilst neglecting the intervening barriers. The relationship between $U(\mathbf{r})$ and $\tilde{U}(\mathbf{r})$ is shown pictorially in Figure 2.2. By exploring

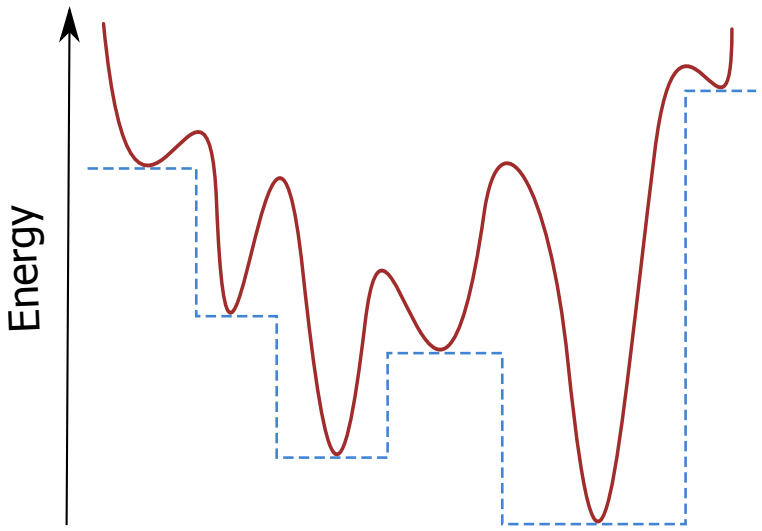


Figure 2.2: Schematic of a one dimensional energy surface $U(\mathbf{r})$, (solid line) and its associated transformed "staircase" surface $\tilde{U}(\mathbf{r})$, (dotted line).

$\tilde{U}(\mathbf{r})$ using Markov Chain Monte Carlo, one can in effect transition (or "hop") between the basins on $U(\mathbf{r})$ without being impeded by the barriers which separate them.

In practice, the basin-hopping method is implemented as follows;

1. Starting from an initial potential energy minimum a with coordinate $\mathbf{r}_a$ and potential energy U_a, a perturbation to the geometry $\delta\mathbf{r}$ is made.

2. The perturbed geometry is locally optimised with respect to the potential energy to a minimum b with coordinate $\mathbf{r}_b$ and potential energy U_b. This is equivalent to applying the transformation in Equation 2.23.

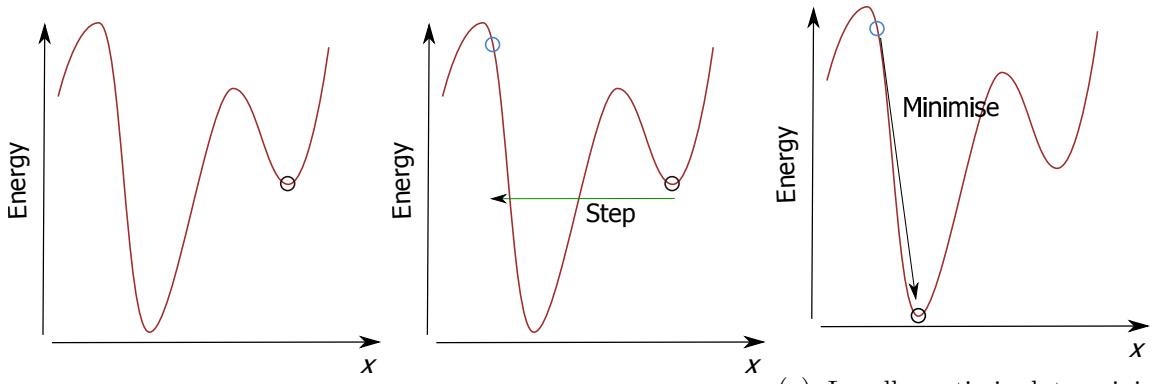
3. The new geometry is accepted or rejected according the Metropolis criterion [7],

$$p(a \to b) = \min\left[1, \exp\left(\frac{-\Delta U}{k_B T}\right)\right] \tag{2.24}$$

where $\Delta U = U_b - U_a$, T is a fictitious "temperature" used to scale the acceptance probability of uphill steps and k_B is the Boltzmann constant.

4. Return to step 1 and repeat. If the new geometry was accepted, replace minimum a with b, otherwise reuse minimum a.

This is repeated until some terminating criterion is reached (typically a maximum number of steps or having successfully located a target structure/energy).



(a) Minimum a with energy U_a (b) A random step $\delta\mathbf{r}$ (c) Locally optimised to minimum b with energy U_b

Figure 2.3: Schematic of a single step in a basin-hopping search of a one dimensional energy landscape.

A few opening observations regarding the method; unlike in regular Monte Carlo simulations, T does not correspond to the temperature of the equilibrium ensemble that is being sampled. Instead, it is a fictitious "temperature" used to control the acceptance probability of moving from a lower energy minimum to a higher energy minimum and has no physical meaning. Also, the size of $\delta\mathbf{r}$ is larger than in regular Monte Carlo simulations, as it is required to move *between* basins, instead of sampling within a basin. Choosing an appropriate value for T, and choosing a set of move classes from which to draw $\delta\mathbf{r}$ is critical to the efficient performance of the method. There is no globally optimum choice, so one must consider the nature of the chemical system at hand, and the topography of its energy landscape. The value of T required for a particular system is probably determined by the topography of its

energy landscape. For a funnelled landscape, moving towards the global minimum correlates with moving to progressively lower energy minima. In this instance, a low value of T would be preferable, because it would direct successive steps towards the bottom of the funnel. In contrast, a frustrated landscape would require a higher value of T, as the method would need to be able to escape from any metastable funnel that it might encounter. The choice of move classes from which to draw random geometry perturbations is usually influenced by the chemical nature of the system. A homoatomic cluster might only require a random displacement of its constituent atoms, whereas a heteroatomic cluster would also benefit from swapping the position of unlike atom types. Systems with constraints, such as a molecule with a network of covalent bonds would benefit from concerted atomic motion, such as rotations of dihedral angles between bonded atoms which preserves bond lengths and bond angles whilst moving side chains and and active sites.

2.3 Landscape Visualisation

For all but the simplest cases, $U(\mathbf{r})$ has many more degrees of freedom than it is possible to visualise conventionally, making it impossible to assess the surface topography directly. One solution to the visualisation problem is to partition the landscape into discrete regions, and then hierarchically cluster these regions according to some similarity measure. This clustering can then be represented as a tree-graph in either two or three dimensions. There are a number of examples of hierarchical clustering methods in the literature, broadly based on either geometry, energetic barriers, or local ergodicity.

In geometrical clustering, regions are clustered according to their structural similarity, which is usually defined by the root-mean-square deviation (RMSd) between them. In this context, regions can either correspond to minima on $U(\mathbf{r})$ [50] or points along a molecular dynamics trajectory [51, 52]. The structures are clustered either by an iterative process, by which each structure is joined to its nearest neighbour until only a single cluster remains [51], or clustering structures that are within a critical distance of one another [50].

2.3.1 Disconnectivity Graphs

When clustering according to energetic barriers, the landscape is partitioned into basins of attraction whereby each point on the landscape $U(\mathbf{r})$, is mapped onto a local minimum α, with co-ordinate, $\mathbf{r}_\alpha$ by a steepest-descent path [53, 45]. Alternatively, the landscape can be partitioned using a lumping approach [54], in which energy thresholds are used to group connected regions below the threshold. The similarity measure used for hierarchical clustering is the barrier energy that separates any two regions. Starting from the energy of the global minimum, U_0, regions are clustered together if they are separated by a barrier with an energy lying in the interval $U_{i+1} - U_i$, where $U_{i+1} = U_i + \Delta U$ and ΔU is the width of the interval. This clustering is repeated until a particular energy threshold, U_t, is reached, or all the minima are clustered together. Such graphs have come to be referred to as disconnectivity graphs [53, 55], and have been used in a number of studies to visualise energy landscapes [34, 53, 55, 56]. Disconnectivity graphs retain both the energies of minima

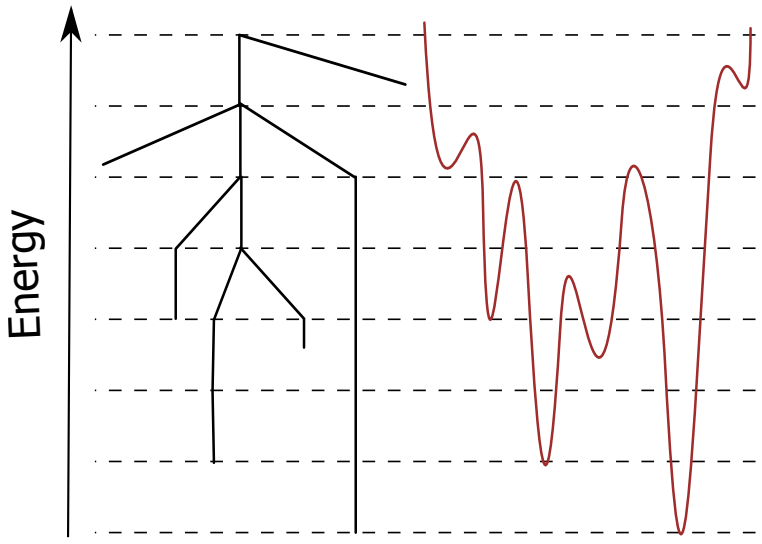


Figure 2.4: Schematic of a one dimensional energy surface $U(\mathbf{r})$, (right) and its associated disconnectivity graph (left).

on $U(\mathbf{r})$, and the barriers that separate them, making them a useful diagnostic in visually assessing the thermodynamic and kinetic behaviour of a system [57, 58]. They can also be used to represent free-energy surfaces by estimating the vibrational entropy of minima and transition states on the landscape from the harmonic superposition approximation [59, 60, 61, 62]. Clustering landscapes by local ergodicity involves partitioning the landscape

into basins about local minima. Equilibration between basins is determined by comparing forward and backward transition rates between states [63] or the time-dependent probability distributions of connected basins [64].

A weakness of the disconnectivity graph method is that it does not retain any structural information on the minima and thus neglects a large portion of the information contained in the energy landscape. Metric disconnectivity graphs capture some of this structure by defining a metric, and then calculating an order parameter from the metric for each minimum of interest on the landscape. The minima can then be plotted along a metric axis perpendicular to the energy axis. Metric information can be included in a number of other ways, such as by changing the colour, or thickness of the nodes and edges [65, 66]. In this thesis, we will refer to metric disconnectivity graphs (MDGs) as those for which the nodes are organised along a metric axis. 2D metric disconnectivity graphs are plotted with the position of the minima on the x-axis defined according to a metric. In 3-D disconnectivity graphs, two metrics are used. The positions of nodes on the disconnectivity graphs are defined as the mean of the metrics for all minima connected to that node. A judicious choice of metric captures overall structural trends in the system, whilst ignoring noisy or irrelevant information.

2.4 Software

2.4.1 PyConnect `https://github.com/lsmeeton/pyconnect`

Disconnectivity graphs and metric disconnectivity graphs presented in this thesis are plotted using PyConnect. The PyConnect package comprises two components: PCA, which calculates the principal components of molecular systems from PATHSAMPLE [21] databases, and PyConnect, which constructs and displays metric disconnectivity graphs. Both PCA and PyConnect were written in Python. The disconnectivity graphs are rendered with MatPlotLib [67], and users can chose to create disconnectivity graphs and metric disconnectivity graphs in 2 or 3 dimensions. PyConnect also provides some cosmetic features, including the ability to label minima, colour minima according to an order parameter or according to their basin of residence. PyConnect can also be used to modify graphs interactively using

the iPython [68] virtual environment.

Plotting Disconnectivity Graphs

In the disconnectivity graphs produced by PyConnect, the position of nodes and minima along the x-axis are determined by algorithms similar to those used in DISCONNECT [69], another program for producing disconnectivity graphs from databases of minima and transition states. PyConnect builds a disconnectivity graph from a database of minima and transitions states as follows:

1. The minima and transition states contained in the database files are read and saved into memory.

2. The database is then pruned;

   i Firstly, the minima and transition states which are not connected to a particular minimum (by default the global minimum) below the energy threshold `FIRST` are discarded from the database.

   ii The database is checked to ensure that it is connected and that the stationary points are physically plausible (i.e. the transition state joining a pair of minima has a higher energy than both of them).

3. Once this database has been built, the minima can be organised into basins, a basin being all the minima which are connected below a particular energy level, and can contain either a single minimum or a number of minima.
   The tree graph structure has now been determined, and the graph can be plotted.

4. For metric disconnectivity graphs, plotting the graph is fairly trivial. Nodes are positioned according to the mean value of the metric for each minima they contain. For disconnectivity graphs, positioning the nodes and edges can be a little trickier, as each node and edge must be placed for maximum clarity and to avoid edge-crossing. This is achieved as follows;

i The x-axis is divided into "columns", with a column for each minimum contained in the graph.

ii Starting from the highest energy level and iterating through to the lowest;

a Nodes are assigned a column for each minima they contain

b Nodes are positioned below their parents, with the largest nodes positioned as close to the centre of their parents, and smaller nodes positioned progressively further out

5. Once the positions of all the nodes are determined, the graph can be plotted using matplotlib.

Part I

Biomolecules

Chapter 3

Visualising the Energy Landscapes of Model Proteins

In this chapter, the analysis of two model protein macromolecules using metric disconnectivity graphs is presented. The chapter also includes a discussion of the wider topic of proteins in energy landscape theory, as well as methodology specific to calculating the order parameters used in the metric disconnectivity graphs.

3.1 Proteins in Energy Landscape Theory

Proteins are linear heteropolymers, responsible for a wide and diverse range of biological processes and physiological functions. The efficacy with which a protein is able to perform its desired function depends critically upon it being able to adopt a very specific three dimensional structure, known as the native state or conformation. For small globular proteins (those composed of $\lesssim 300$ monomers), the structure of the native state is determined solely by the order in which the protein monomer units appear along the length of the chain. Though possible in principle, predicting the native conformation of a given protein from the ordering of its monomer units is an extremely taxing problem, and remains the focus of much current research.

In this section two concepts are examined in brief. Firstly, an overview of how energy landscape theory can be used to predict native protein conformations and protein folding

pathways is given. Secondly, the BLN and Gō model proteins are introduced as a useful tool in understanding and predicting such folding behaviour. Proteins are responsible for a wide range of biological processes, and perform a variety of physiological functions. Protein molecules such as the keratin in hair and nails and collagen in connective tissue can be used to confer rigidity to biological tissue. Protein function can also include transport, storage and catalysis, examples of which include haemoglobin (oxygen storage), cytochrome *c* (electron transport) and lactase (enzyme required for the digestion of lactose in milk) [70].

Proteins are composed of *amino acid* monomers covalently linked to one another by *peptide* bonds. The monomers are formed of a common backbone and a side chain which determines the identity of the amino acid, and takes the general form $NH_2C_\alpha RHC'O_2H$, where R is the side group (see Figure 3.1a). The side chain can be one of approximately 20 different organic chemical species, spanning from a single hydrogen atom in the case of glycine to large aromatic side chains such as tryptophan, which has a mass of 204 Da (where 1 Da $\equiv$ 1 unified atomic mass unit). The peptide bond has a partial double bond character which inhibits

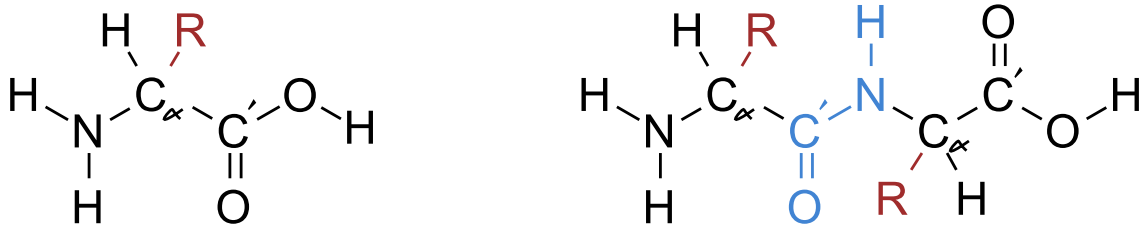


Figure 3.1: The general form of an amino acid (a) and a peptide bond (b). The side chains are highlighted in red and the peptide bond is highlighted in blue.

rotation about it (there is a resonance between two extreme structures, one which allows free rotation, and another which restricts rotation but has a large dipole moment [71]). Due to these resonances, and reasonably rigid bond lengths and bond angles, the only degrees of freedom in a peptide bond are rotations about C_α–N and C_α–C bonds, represented by the dihedral angles ϕ and ψ. This results in the conformation of the protein backbone requiring just two dihedral angles per residue to describe it [72].

The structure of a protein molecule is described at four hierarchical levels; Primary, Secondary, Tertiary and Quaternary.

Primary structure describes the order of the amino acid residues along the length of the protein.

Secondary structure describes the local spatial configuration of the protein. A common secondary structure is the α–helix, where every backbone N–H group donates a proton to the backbone C=O group four residues earlier to form a hydrogen bond. The α–helix is particularly stable because it is maximally hydrogen bonded [70]. Another common structural motif is the β–sheet, in which two or more sections of the polypeptide chain link in a regular manner, with hydrogen bonds forming between the N-H and C=O groups of the protein backbone [73].

Tertiary structure is the overall 3D conformation of the protein, detailing how secondary structural elements interact with each other. The information present in the primary structure is distributed both locally and non locally, and secondary structures is as much a result of tertiary structure as it is responsible for it [74].

Finally, *quaternary structure* describes the aggregation of individual proteins to form a larger, more complex macromolecule. Examples of quaternary protein complexes are haemoglobin, which is an aggregate of four myoglobin-like proteins, and is found in the blood of vertebrates to perform oxygen transport [70].

In order for a protein to perform its desired function, it must fold into the tertiary or quaternary structure corresponding to its native conformation. For small globular proteins all the information required to reach the native state is present in the primary structure of the protein, and no external help is required (note that for larger proteins, spontaneous folding can be harder to achieve, and the help of molecular chaperones are required) [75, 76, 77]. One can now make a comment on the potential difficulty faced by a protein undergoing a folding process into its native state. As the length of the protein chain increases, the number of possible confirmations the chain can adopt increases exponentially (for a chain of 100 residues there are $\sim 10^{70}$ possible conformations). Assuming that the protein searches its configuration space randomly, and at the quickest rate possible ($\sim 10^{-11}$ seconds) it would take on average about 10^{52} years to find the native conformation. Of course a protein can fold into its native state much quicker than the lifetime of the universe, and this apparent

paradox was first identified by Levinthal [78, 11]. Levinthal's paradox assumes that in order to find the native state, the protein must randomly search its entire configuration space, which is equivalent to a trajectory across a flat energy landscape. However, if a small energetic bias towards the native state is introduced, the required search time rapidly reduces to biological time scales [79]. This resolution is referred to as the thermodynamic hypothesis, which postulates that the native conformation of a protein corresponds to the global energy minimum. This reduces the determination of the native state of a given protein to a global optimisation problem, and thus can be tackled using energy landscape theory [75].

Of course, global optimisation and energy landscape exploration are by no means trivial, especially for a larger, flexible molecule like a protein with many effective degrees of freedom and competing intra-molecular interactions. Native proteins are typically only 5–10 kcal mol^{-1} more stable than their denatured state, so it is clear that no single type of interaction can be neglected in folding and structure predictions. Electrostatic interactions occur due to charged residues, though they are not likely to dominate as they are concentrated in high-dielectric regions on the protein surface. Hydrogen bonding is important because all possible hydrogen bonding interactions are generally satisfied in native structures, and are key in all secondary structures. Van der Waals effects must also be considered, as when in the native state, the protein structure is tightly packed. Long-range, water mediated interactions also play a role, and require a detailed understanding of the solvent-residue interactions, which can add computational cost and extra degrees of freedom [80].

As a result, it is often preferable to study a simplified or toy protein model, instead of attempting to faithfully model the specifics of a real protein (though of course that too has its merits). Using an appropriately designed toy model, one can explore a particular aspect of protein physics such as a folding behaviour or the effect of differing primary structure on native state tertiary conformation, but with considerably less computational effort. One should note that a 'third way' exists between all-atom and toy model proteins; the coarse grain protein. Coarse grained protein models such as MARTINI [81] attempt to replicate protein structure by using 'beads' which represent specific combinations of approximately 4 heavy atoms represent. In principle, a coarse grained model should be able to approx-

imate some desired physical property of a real biological system, but at a fraction of the computational cost.

3.1.1 BLN and Gō Model Proteins

The BLN and Gō models are toy proteins which have been used extensively to study protein folding behaviour. The BLN model is a coarse-grained, off–lattice protein model in which each protein residue is represented by one of three types of bead: hydropho**B**ic, hydrophi**L**ic or **N**eutral [82, 83]. In this thesis, a version of the BLN potential in which the inter-residue distances and angles are restrained with stiff springs is used [84]. The beads interact with one another according to

$$\begin{aligned} U_{BLN}(\mathbf{r}) =& \frac{1}{2}K_d \sum_{i=1}^{N-1} (d_{i,i+1} - d_e)^2 + \frac{1}{2}K_\theta \sum_{i=1}^{N-2} (\theta_i - \theta_e)^2 \\ &+ \varepsilon \sum_{i=1}^{N-3} \{A_i (1 + \cos\phi_i) + B_i (1 + 3\cos\phi_i)\} \\ &+ 4\varepsilon \sum_{i=1}^{N-2} \sum_{j=i+2}^{N} C_{ij} \left\{ \left(\frac{\sigma}{d_{ij}}\right)^{12} + D_{ij} \left(\frac{\sigma}{d_{ij}}\right)^{6} \right\} \end{aligned} \tag{3.1}$$

where d_{ij} is the distance between two beads i and j. The first term is a harmonic bond restraint with $K_d = 231.2\epsilon\sigma^{-2}$ and $d_e = \sigma$. The second term represents a harmonic angle constraint the $K_\theta = 20$ rad^{-2} and $\theta_\varepsilon = 1.8326$ rad. The third term takes into account torsional angles along the chain, and is defined by four consecutive beads. If two or more beads are N, then $A = 0$ and $B = 0.2$, else $A = B = 1.2$. The fourth term represents long range, water mediated hydrophobic interactions between non-bonded pairs. If both beads are B, then $C = D = 1$. If one residue is L and the other is L or B, then $C = \frac{2}{3}$ and $D = -1$. If either residue is N, then $C = 1$ and $D = 0$ [84].

A number of bead sequences of differing length have been proposed and studied for the BLN model, including 46 and 58 bead polymers [85, 86, 87, 84]. In this thesis, we consider only the BLN-69 model, which consists of 69 beads with the sequence [88]

$$\mathrm{B_9N_3(LB)_4N_3B_9N_3(LB)_4N_3B_9N_3(LB)_5L} \tag{3.2}$$

BLN-69 has been designed to exhibit a frustrated energy landscape, with a 6-strand β-barrel structure as its global minimum. The model has been shown to have a number of low–energy β-barrel-like structures, which differ from the global minimum by a chain slip along the length of the barrel [89], but which are separated by large barriers. As an aside, one should note that this is clearly not a good model for protein folding, as the system kinetics are frustrated and do not exhibit a single, structure seeking super-basin which acts to funnel folding trajectories into a well-defined native state. Such frustration is absent when considering the "Gō" version of the model (Gō-69), where attractive interactions between pairs of residues that are not in contact in the native state (i.e. the global minimum) are neglected [90, 91].

3.2 Collective Variables

In this section, the challenge of computing collective variables is addressed. The purpose of computing collective variables is to find some small number of coordinates which are somehow representative of the molecular system of interest as a whole. This is not an easy task, with methods typically focussing on one of two broad themes. The first involves leveraging ones chemical intuition of a system to gain insight. For example, if the system of interest adopted a crystalline structure at low energy, an appropriate collective variable might be the local bond order parameters [92]. Alternatively, methods which determine collective variables whilst requiring minimal *a priori* knowledge of the system's chemistry can be implemented. Methods from the field of machine learning are often good candidates.

In the following section, four methods for computing collective variables are described, each of which was used to investigate the BLN-69 and Gō-69 model proteins. The native contact and RMSd metrics both rely on some prior knowledge of the system (that proteins are structure seeking systems and thus structural similarity to the global minimum structure is a good progress variable). The principal component and isomap metrics are part of the latter class of collective variables. They are both machine learning methods which attempt to determine global structural properties of the protein from the distribution of minima in

configuration space.

3.2.1 Native Contact Metric

The native contact metric evaluates for each minimum, α, the ratio N_α/N_{NC}, where N_{NC} is the number of native contact pairs and N_α is the number of contact pairs in minimum α that are also present in the native conformation. Here, contacts are defined as those beads which are within 1.167σ of each other, excluding pairs that are within three beads of each other in the peptide sequence [89]. Hydrogen bonding is important in protein folding,

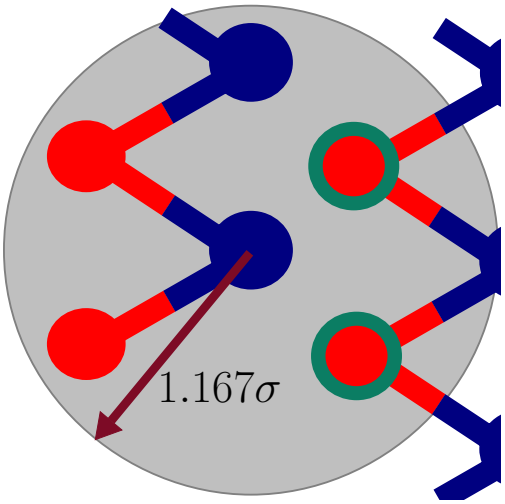


Figure 3.2: Cartoon of the native contacts in a protein chain. For the central blue peptide bead, native contacts are all those peptide beads in the chain which are within 1.167σ (i.e. within the grey circle) excluding pairs that are within three beads of each other in the peptide sequence. The peptide beads in the cartoon which are native contacts are highlighted in green.

and native contact analysis can provide a useful analogy for coarse grained protein models. N_α/N_{NC} is commonly used as a progress variable in computational studies of protein folding to distinguish between the different degrees of partially folded protein [93].

3.2.2 RMSd Metric

The RMSd measures the distance between the conformation of minimum α and β, $\mathbf{r}_\alpha$ and $\mathbf{r}_\beta$ respectively, according to

$$\mathrm{RMSd}\,(\alpha,\beta) = \sqrt{|\mathbf{r}_\alpha - \mathbf{r}_\beta|^2} \tag{3.3}$$

Invariance under global translations and rotations is implicit if structures are represented in internal co-ordinates. When working in cartesian co-ordinates, the Kabsch algorithm [94] was used to align structures to minimise RMSd (α, β). In the RMSd metric, RMSd (α, β) is calculated between the conformation of each minimum and the conformation of the global minimum, $\mathbf{r}_{GM}$.

3.2.3 Principal Component Metric

The principal component metric is based on principal component analysis (PCA), a statistical procedure used to analyse large, high-dimensional data sets, which is commonly used in dimensional reduction and, or when the relevant degrees of freedom in a data set are not clear [95]. For a data set with an initial basis $\{\mathbf{e}_i\}$, the principal components form a basis set $\{\mathbf{q}_i\}$ which is related to $\{\mathbf{e}_i\}$ by a *linear* transformation. The basis of principal components are incident to the axes of greatest sample variance of the data set, with the principal component lying along the the axis of greatest sample variance referred to as 1^{st} principal component (Q_1). The principal component which lies along the axis of second greatest sample variance (orthogonal to the first) is the 2^{nd} principal component (Q_2) and so on [96]. We performed PCA on the set of N_{sp} stable configurations $\{\mathbf{r}_\alpha\}_{U_t}$, where $\{\mathbf{r}_\alpha\}_{U_t}$ are

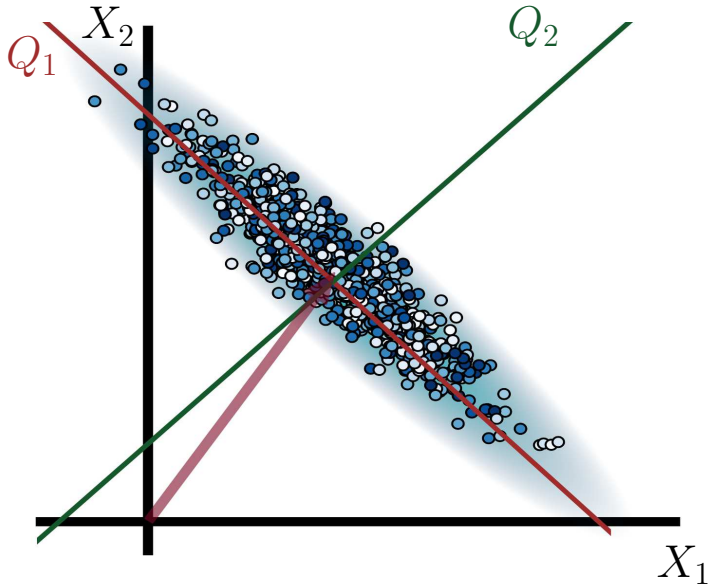


Figure 3.3: Cartoon of the principal components of a two dimensional data set. The data is represented in an initial basis set $\{X_1, X_2\}$. The principal components Q_1 and Q_2 lie along the first and second axes of greatest sample variance. Before principal component analysis is performed, the data set needs to be transformed so that the mean sits at the origin.

all local minima connected to the global minimum below a certain threshold energy U_t. As PCA is a linear method, the principal components can be sensitive to the initial basis set in which the data set is expressed, i.e. if a data set is expressed in two different basis unrelated by a linear transformation, then the principal components will be different. For the results presented, the initial basis sets employed were the basis of external cartesian coordinates, $\{\mathbf{e}_i\}$, and an internal basis set of dihedral angles, $\{\psi_i\}$. For the BLN and Gō model proteins, all the beads were used for the PCA. For the cyclic peptides, only the heavy atoms in the peptide backbone were considered.

Before PCA can be performed, $\{\mathbf{r}_\alpha\}_{U_t}$ must be pre-processed, with the exact procedure depending the basis set under consideration. Due to the cyclicity of angular coordinates, they cannot be used directly as a basis to compute principal components. To remove the periodicity of $\{\psi_i\}$, we used as a basis set the sines and cosines of the internal dihedrals, $\{\cos\psi_i, \sin\psi_i\}$ [97]. This transformed representation of the backbone dihedrals *is* linear, but requires that the number of basis vectors be doubled. For the external cartesian coordinates, spurious sample variance must be removed by ensuring that each structure in $\{\mathbf{r}_\alpha\}_{U_t}$ is translated and orientated to be as close to one another as possible. The rotational and translational invariance of $\{\mathbf{r}_\alpha\}_{U_t}$ was enforced by implementing McLachlan's best fit procedure [98], which works as follows;

1. A reference configuration defined as the ensemble average, $\langle\mathbf{r}\rangle$ of $\{\mathbf{r}_\alpha\}_{U_t}$ was calculated, where $\{\mathbf{r}_\alpha\}_{U_t}$ is the set of N_{sp} minima of interest, and where each configuration in $\{\mathbf{r}_\alpha\}_{U_t}$ has its centroid centred on the origin.

2. Define a new set $\{\mathbf{r}_\alpha\}_{U_t}$, rotate each configuration about its origin to be as close to $\langle\mathbf{r}\rangle$ as possible using the Kabsch algorithm, and thus minimise

$$s = \frac{1}{2}\sum_{\alpha=1}^{N_{sp}} (\mathbf{r}_\alpha - \langle\mathbf{r}\rangle)^2 \tag{3.4}$$

3. Replace $\{\mathbf{r}_\alpha\}_{U_t}$ with $\{\mathbf{r}'_\alpha\}_{U_t}$.

4. Repeat steps 1–3 until the ensemble average converges to some threshold criterion.

For the results presented, we used the threshold criterion defined by Komatsuzaki *et al.* [65], $s \leq 10^{-8}$.

Once $\{\mathbf{r}_\alpha\}_{U_t}$ has been suitably pre-processed it can undergo PCA. Hereafter, whether discussing cartesian or internal coordinates, we define $\{\mathbf{r}_\alpha\}_{U_t}$ as the translation-free, rotation-free set of configurations, and $\{\mathbf{q}_i\}$ as the basis set, where i is the coordinate index. To perform PCA, we begin with defining the $3N \times N_{sp}$ mean-centred configuration matrix, $\mathbf{R}$,

$$\mathbf{R} = \begin{pmatrix} & & & & \\ \mathbf{r}_1 & \cdots & \mathbf{r}_\alpha & \cdots & \mathbf{r}_N \\ & & & & \end{pmatrix} \tag{3.5}$$

where each column of $\mathbf{R}$ is a $3N$ dimensional vector corresponding to a stable configuration in the set $\{\mathbf{r}_\alpha\}_{U_t}$. The PCs are the eigenvectors of the $3N \times 3N$ co-variance matrix, $\mathbf{C}$,

$$\mathbf{C} = \mathbf{R}^\dagger \mathbf{R} \tag{3.6}$$

and are thus the basis set $\{\mathbf{Q}_i\}$, where i is the co-ordinate index, in which $\mathbf{C}$ is diagonalised. The PCs are calculated using the *singular value decomposition* method, which states that a $3N \times N_{sp}$ configuration matrix, $\mathbf{R}$, can be written as the product

$$\mathbf{R} = \mathbf{W}\mathbf{S}\mathbf{V}^\dagger \tag{3.7}$$

where $\mathbf{W}$ is the $3N \times 3N$ matrix:

$$\mathbf{W} = \begin{pmatrix} & & & & \\ \mathbf{Q}_1 & \cdots & \mathbf{Q}_i & \cdots & \mathbf{Q}_{3N} \\ & & & & \end{pmatrix} \tag{3.8}$$

The columns of $\mathbf{W}$ are the PCs of the covariance matrix, $\mathbf{C}$, and $\mathbf{S}$ is a $3N \times N_{sp}$ matrix with diagonal elements, S_{ii}, where (dropping the double index for clarity) S_i^2 is the variance associated with $\mathbf{Q}_i$. The PCs are ordered so that $\mathbf{Q}_1$ is the 1^{st} principal component and

has the greatest variance, $\mathbf{Q}_2$ is the second principal component and has the second greatest variance, and so on. To compute the value of the ith principal component metric for a particular minimum α with coordinates $\mathbf{r}_\alpha$, one can transform the basis of $\mathbf{r}_\alpha$ into the basis of principal components, and take the magnitude of $\mathbf{Q}_i$ as the ith principal component metric for minimum α.

A nice feature of PCA is that because the principal components are simply a linear transformation of the original basis set of $\{\mathbf{r}_\alpha\}_{U_t}$, the principal components can be projected back onto the original basis set. This enables one to visualise the PCs of a given $\{\mathbf{r}_\alpha\}_{U_t}$ by choosing a reference structure, $\mathbf{r}_{ref}$, and adding the $\mathbf{Q}_i$ of interest to it

$$\mathbf{r}_\lambda = \mathbf{r}_{ref} + \lambda \mathbf{Q}_i \tag{3.9}$$

where λ is a progress variable. For the cartesian basis set, this means that one can see the direction in configuration space in which the ith principal component $\mathbf{Q}_i$ is incident.

3.2.4 Isomap Metric

The Isomap metric is based on the Isomap algorithm [99, 100], a nonparametric, nonlinear dimensionality reduction technique. The aim of the Isomap algorithm is to define a low-dimensional embedding that as accurately as possible preserves geodesic distances between all pairs of points in the data cloud. The geodesic distance between a pair of points that lie on a manifold is the length of the shortest path between them that lies along that manifold. Isomap assumes that such a low-dimensional manifold exists, and that its shape can be estimated from the distribution of points in the data cloud. The Isomap algorithm approximates the geodesic distance between a given pair of points on the manifold by calculating the shortest possible path between them that can be found by stepping from one point to its neighbour.

In Figure 3.4 an illustration of the Isomap algorithm is shown, using the three dimensional 'Swiss Roll' data set, represented as the set of data points plotted on the left-hand graph. From visual inspection one can observe that the data is not distributed extensively through-

out the input space, but in fact lies on a two dimensional 'Swiss Roll' manifold. In the left hand graph two arbitrary data points have been circled in black, and the three dimensional Euclidean distance between the two is shown as a dotted line, which may not as accurately show the similarity between the highlighted point as the geodesic along the 'Swiss Roll' manifold, shown in blue. In the middle graph, a graph of nearest neighbours has been constructed from the data points, and approximation to the geodesic has been computed, shown in red. In the final graph, the two dimensional 'Swiss Roll' has been unfolded along the two dimensional embedding recovered by Isomap. We applied Isomap to the set of N_{sp} stable

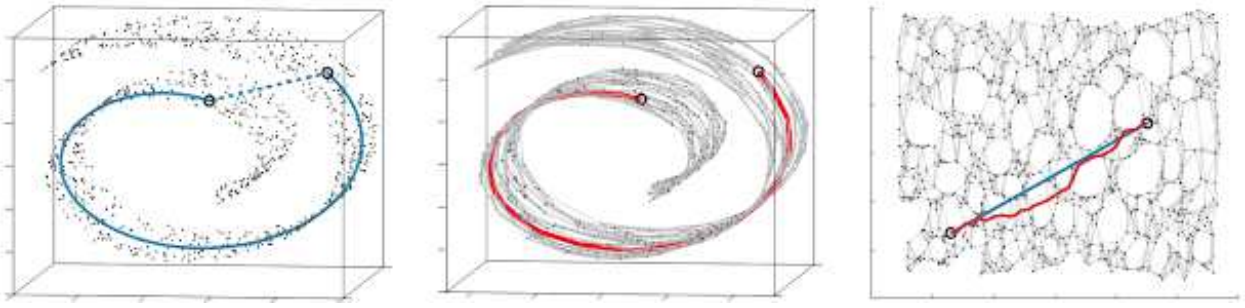


Figure 3.4: An illustration of the Isomap algorithm using the three dimensional 'Swiss Roll' data set, taken from [99].

configurations $\{\mathbf{r}_\alpha\}_{U_t}$ using the Isomap implementation in the scikit-learn machine learning package [101]. As with the principal component metric, rotational and translational invariance of $\{\mathbf{r}_\alpha\}_{U_t}$ was enforced by implementing McLachlan's best fit procedure.
The Isomap algorithm consists of three steps:

1. A weighted graph, G, of $\{\mathbf{r}_\alpha\}_{U_t}$ is built, where each conformation is a node and where the k nearest-neighbours of each conformation α are joined by an edge with weight $d_{\alpha\beta}$. Isomap has been shown to be fairly robust to the choice of k [100], and in this study we took $k = 15$.

2. The shortest path between each conformation through the graph G is determined and a distance matrix, $\mathbf{D}$, is computed, where $D_{\alpha\beta} = \min\{d_{\alpha\beta}, d_{\alpha\gamma} + d_{\gamma\beta}\}$ for $\gamma = 1, \ldots, N_{sp}$. The elements $D_{\alpha\beta}$, are the approximate geodesics between conformations α and β.

3. Classical multidimensional scaling is applied to the matrix D, producing a low-dimensional embedding of the conformational coordinates that best preserves geodesic distances on

the manifold.

The ith isomap metric corresponds to the ith embedded dimension of the low dimensional manifold of $\{\mathbf{r}_\alpha\}_{U_t}$. Unlike with PCA, it is currently not possible to visualise an embedded dimension in the same way that one can project a principal component onto a reference structure (see Equation 3.9).

3.3 Visualising the Energy Landscapes of BLN-69 and Gō-69

In this section, metric disconnectivity graph analysis is presented for two systems, the BLN-69 model protein and its Gō equivalent. For both systems, the database of minima and transition states was compiled with discrete path sampling as implemented in the PATHSAMPLE program [21].

3.3.1 BLN-69

For BLN-69, a database containing 141835 minima and 173692 transition states was used in the study. The first three cartesian and dihedral principal components for the sets, $\{\mathbf{r}_\alpha\}_{U_t}$, connected to the global minimum below energy U_t, where $-95.0\varepsilon \geq U_t \geq -98.0\varepsilon$ are shown in Table 3.1. For the cartesian PCs, Q_1 captures significantly more of the variance

U_t/ε	N_{sp}	Cartesian PCA			Dihedral PCA		
		S_1^{cart}	S_2^{cart}	S_3^{cart}	S_1^{di}	S_2^{di}	S_3^{di}
−95.0	6891	25.0	9.2	8.0	12.3	8.5	8.2
−95.5	5973	25.7	8.9	8.1	12.5	8.6	8.4
−96.0	5135	26.0	8.7	8.2	12.3	8.9	8.6
−96.5	4353	27.4	8.5	8.2	12.7	9.2	9.8
−97.0	1611	37.0	10.0	7.9	12.8	10.6	9.7
−97.5	561	21.2	19.0	10.8	15.1	13.7	11.9
−98.0	409	25.8	16.6	10.8	15.6	14.1	12.0

Table 3.1: The variance captured by the first three principal components in cartesian, S_i^{cart}, and dihedral, S_i^{di}, bases of the N_{sp} structures in the sub-level sets of minima below threshold energy U_t for BLN–69.

than Q_2 for all data sets considered. Q_1 for the sub-level set of minima connected to the

global minimum below $U_t = -97.0\varepsilon$, $\{\mathbf{r}_\alpha\}_{U_t=-97.0\varepsilon}$ has the largest fractional variance and therefore this threshold was selected for all disconnectivity graphs. The dihedral PCs have a more uniform variance distribution than the cartesian PCs, with $S_1^{di} \approx \frac{1}{2}S_1^{cart}$, for all $\{\mathbf{r}_\alpha\}_{U_t}$ considered in both BLN-69 and Gō-69. The dihedral PCs are thus not appropriate metrics for studying these systems, and have not been used to create metric disconnectivity graphs. The set of minima where $U_t = -97.0\varepsilon$ is represented as a disconnectivity graph in Figure 3.5. Figure 3.6 shows the low energy minima labelled a–e in Figure 3.5. Minima b-e are all structurally similar to one another with each adopting compact β-barrel geometries and differing from global minimum a by either a chain-slip, chain-reptation or twist in the turn regions, with further details given in Table 3.2. The native contact metric (Figure

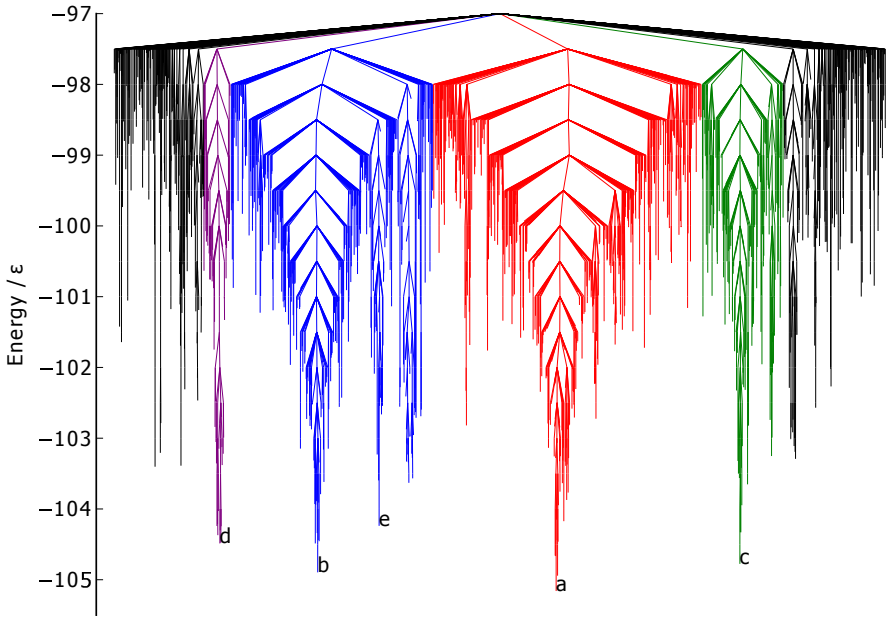
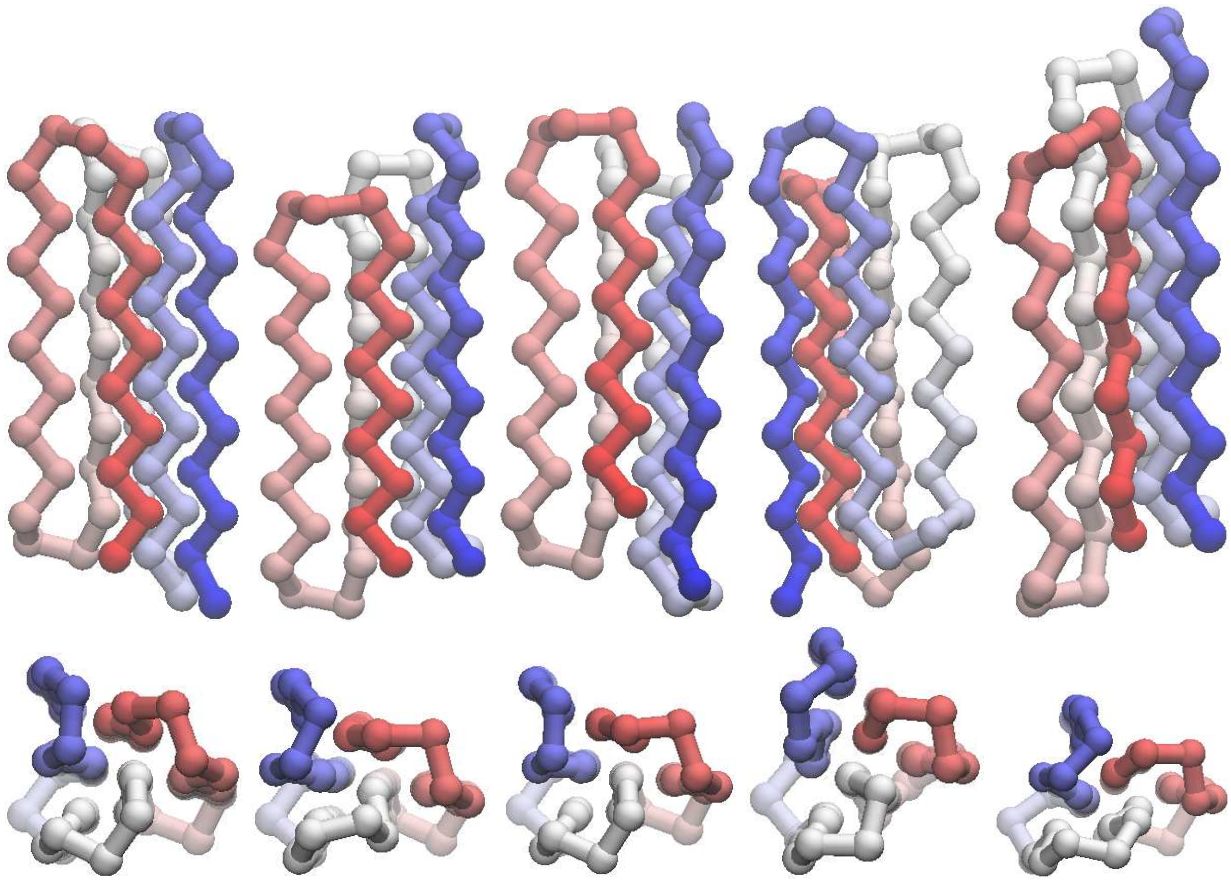


Figure 3.5: Disconnectivity graph of BLN-69, $U_t = -97.0\varepsilon$, $N_{sp} = 1611$. The colour scheme is chosen to distinguish between energetic funnels. Labelled minima correspond to the global minimum and low energy minima separated from the global minimum and one another by large kinetic barriers and are shown in Figure 3.6

3.7) splits the two largest funnels, with each having distinct fractions of native contacts (the mean fraction of native contacts for the funnels containing minima a and b is 0.91 and 0.81 respectively). Though the native contact metric differentiates between kinetically separated minima, it does not differentiate according to their energies. There are a number of unstable,



(a) -105.16 ε (b) -104.90 ε (c) -104.78 ε (d) -104.49 ε (e) -104.24 ε

Figure 3.6: Structures of the minima labelled in Figure 3.5, corresponding to the global minimum, Figure 3.6a and low energy minima separated from the global minimum and one another by large kinetic barriers, Figures 3.6b–3.6e. Energetic and structural details are provided in Table 3.2. The beads are coloured from red to blue (N-terminus to C-terminus).

Minimum	$\Delta U/\varepsilon$	N_α/N_{NC}	RMSd / σ	$\Delta Q_1/\sigma$	$\Delta Q_2/\sigma$	Defect
b	0.26	0.85	0.41	-5.32	-1.25	Chain-slip
c	0.38	0.87	0.21	-0.02	-0.76	Reptation
d	0.67	0.86	0.40	-2.73	-3.51	Double Chain-slip
e	0.92	0.77	0.54	-6.54	-2.61	Twist

Table 3.2: Energy above the global minimum, ΔU, fraction of native contacts, N_α/N_{NC}, RMSd from the global minimum and difference in Q_1 and Q_2 from the global minimum ΔQ_1 and ΔQ_2, respectively, for minima *b*-*e*.

high energy minima with energetically unfavourable turns in the flexible N bead regions, but otherwise with almost all native contacts satisfied. Minimum a by definition satisfies all native contacts. Minima *b*-*d* are all very similar according to this metric, with each satisfying $\approx$ 85% of possible native contacts, in spite of their geometries being relatively dissimilar. The RMSd metric (Figure 3.8) is capable of distinguishing between the different major funnels on the surface, with each having its own mean value of the metric (mean

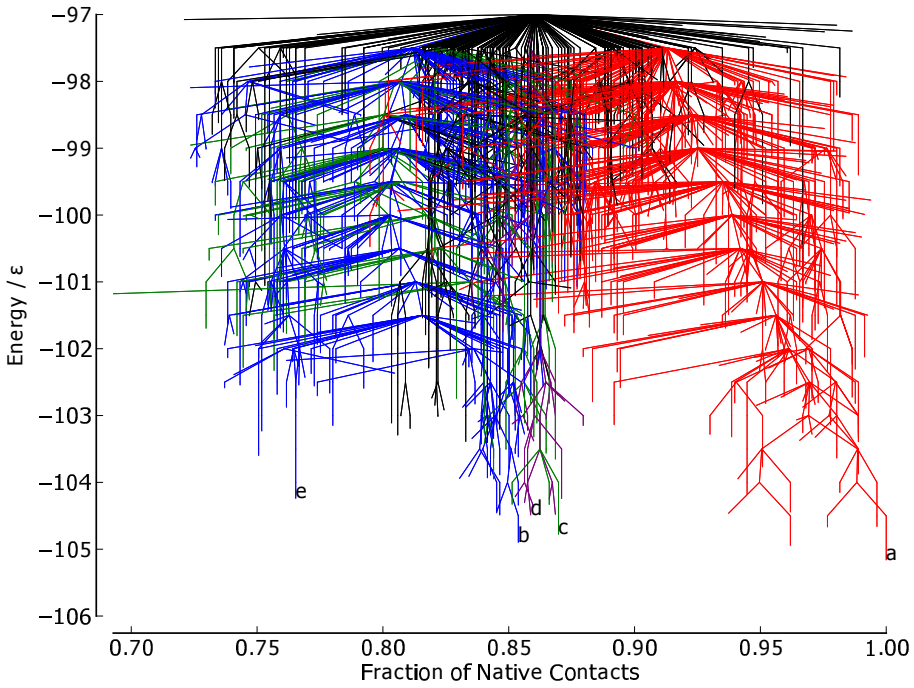


Figure 3.7: Metric disconnectivity graph of BLN-69, $U_t = -97.0\varepsilon$, $N_{sp} = 1611$, with fraction of native contacts used as an order parameter. The colour scheme and labels are as used in Figure 3.5.

RMSd for the funnels containing minima a, c and b 0.24σ, 0.31σ and 0.45σ, respectively). There is also some relation to the minima energy in the green and red funnels, where lower energy corresponds to RMSd metric values closer to 0. The RMSd metric differentiates the basin minima into four groups, with minimum c being most similar to the global minimum, which as expected from a visual inspection of the structures. The Q_1 metric (Figure 3.9) splits the blue funnel (mean 3.29σ) from the red and green funnels, which sit on top of one another (mean 1.59σ and 1.60σ, respectively). The purple funnel lies at the boundary of the two, with a mean of -0.86σ. Minima a and c have almost identical values of Q_1, whilst minima d, b and e have increasingly dissimilar values. Given that Q_1 corresponds to a chain-slip between the C and N termini, and that minima d, b and e have chains that have shifted relative to the global minimum in the same direction, it gives confidence that PCA is capable of identifying structural features of the energy landscape. The Q_2 metric (Figure 3.10) does not reveal any obvious correlation between structure and energetics or kinetics, with no distinction made between the funnels and with the points reasonably evenly distributed along the order parameter. Thus, this PC corresponds to variations within all of

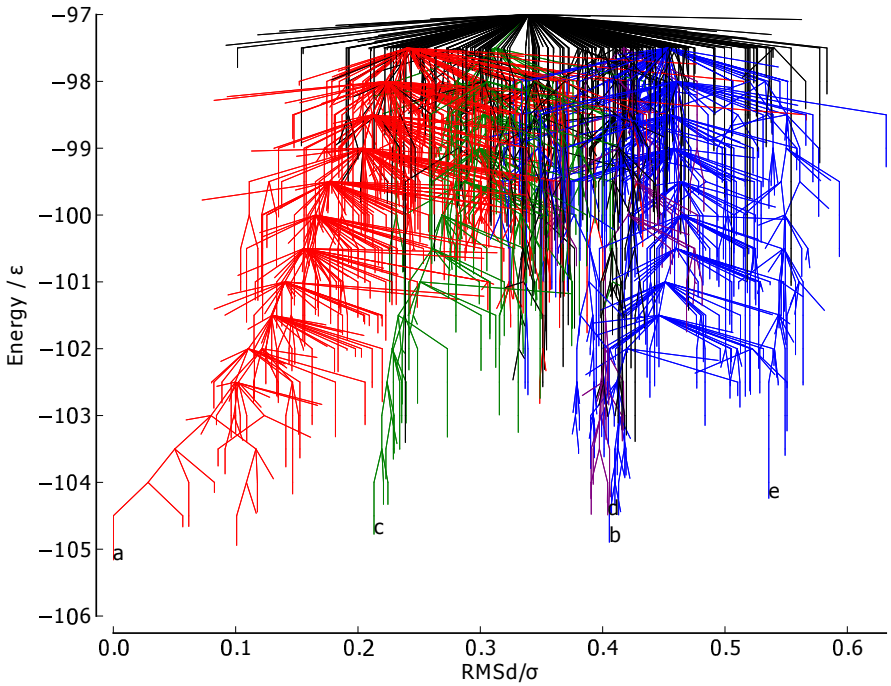


Figure 3.8: Metric disconnectivity graph of BLN-69, $U_t = -97.0\varepsilon$, $N_{sp} = 1611$, with RMSd of each structure from the global minimum used as an order parameter in units of σ. The colour scheme and labels are as used in Figure 3.5.

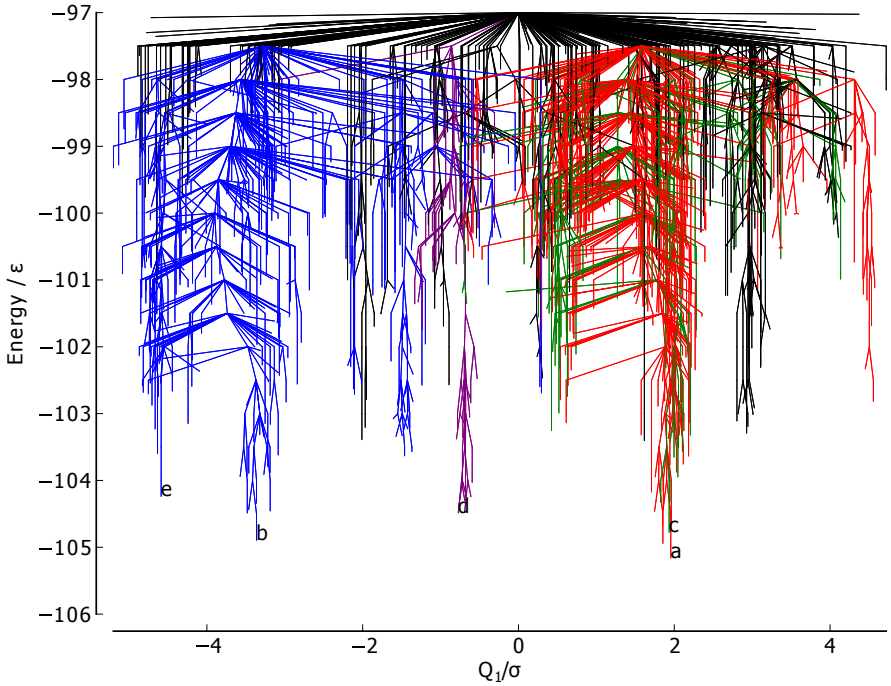


Figure 3.9: Metric disconnectivity graph of BLN-69, $U_t = -97.0\varepsilon$, $N_{sp} = 1611$, with the first principal component of $\{\mathbf{r}_\alpha\}_{U_t=-97.0\varepsilon}$ used as an order parameter in units of σ. The colour scheme and labels are as used in Figure 3.5.

the funnels rather than structural differences between the funnels. The progression of Q_1 of

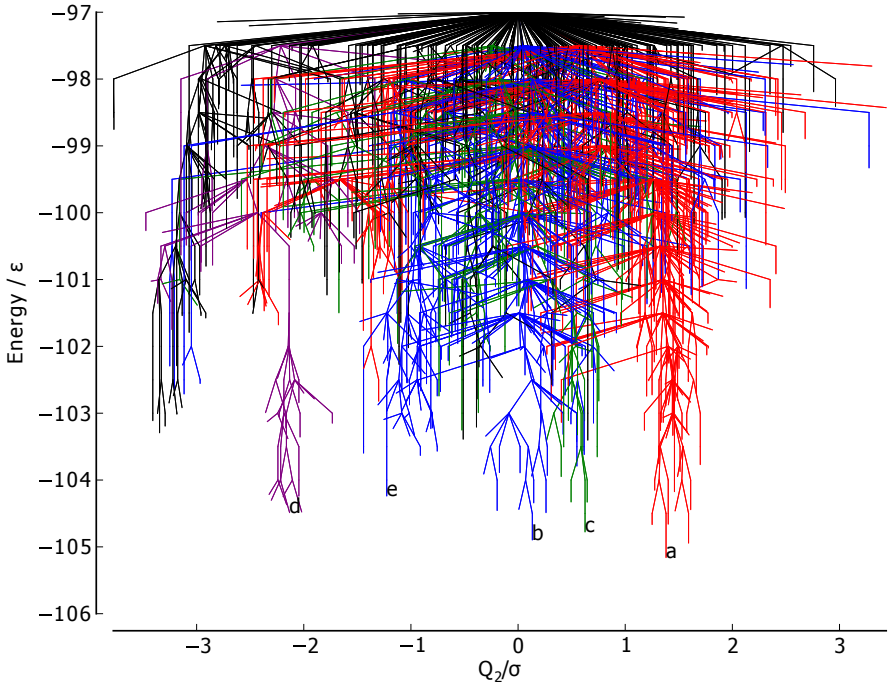
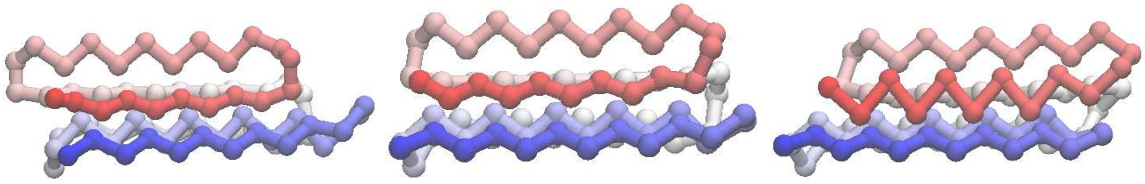


Figure 3.10: Metric disconnectivity graph of BLN-69, $U_t = -97.0\varepsilon$, $N_{sp} = 1611$, with second principal component of $\{\mathbf{r}_\alpha\}_{U_t=-97.0\varepsilon}$, Q_2, used as an order parameter in units of σ. The colour scheme and labels are as used in Figure 3.5.

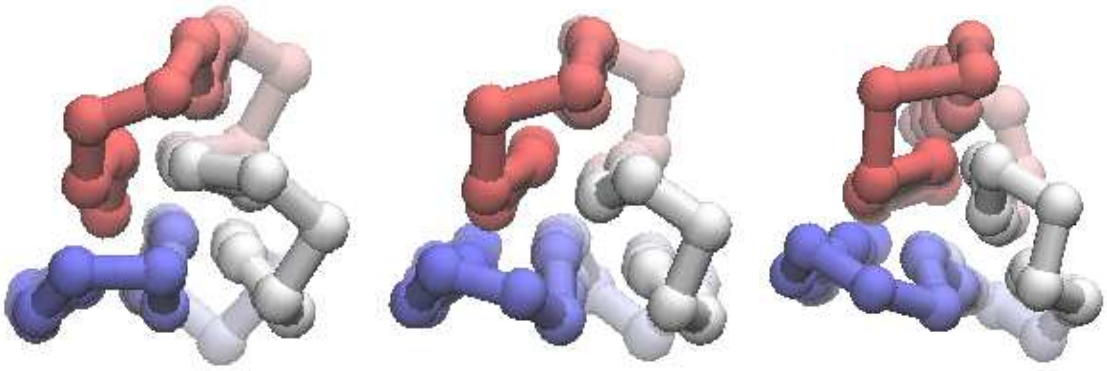
$\{\mathbf{r}_\alpha\}_{U_t=-97.0\varepsilon}$ from $\lambda = -5.2\sigma$ to $\lambda = 4.7\sigma$ (Figure 3.11) corresponds to a chain-slip between the C and N termini. The progression of Q_2 of $\{\mathbf{r}_\alpha\}_{U_t=-97.0\varepsilon}$ from $\lambda = -3.7\sigma$ to $\lambda = 3.4\sigma$ (Figure 3.12) corresponds to a twisting of the internal chain sequences. The 1^{st} embedded



(a) $Q_1 = -5.2\sigma$ (b) $Q_1 = 0.0\sigma$ (c) $Q_1 = 4.7\sigma$

Figure 3.11: Different values of Q_1 for $U_t = -97.0\varepsilon$ projected onto the structure of the global minimum of BLN-69. For the global minimum, $Q_1 = 1.96\sigma$. The beads are coloured from red to blue (N-terminus to C-terminus). An animated version of this projection is available as Supporting Information in [102].

dimension of the isomap metric (Figure 3.13) clearly differentiates between all the coloured funnels on the landscape. The mean value of the blue, purple, red and green funnels are -10.23σ, -2.96σ, 4.12σ and 8.80σ, respectively. The structure of the graph is similar to the Q_1 graph (Figure 3.9), with the order of the coloured funnels and labelled low energy minima



(a) $Q_2 = -3.7\sigma$ (b) $Q_2 = 0.0\sigma$ (c) $Q_2 = 3.4\sigma$

Figure 3.12: Different values of Q_2 for $U_t = -97.0\varepsilon$ projected onto the structure of the global minimum of BLN-69. For the global minimum, $Q_2 = -5.5\sigma$. The beads are coloured from red to blue (N-terminus to C-terminus). An animated version of this projection is available as Supporting Information in [102].

along the metric axis matching. The agreement between these two metrics suggests that the 1^{st} embedded dimension of the isomap metric is fairly linear, and that, as with the Q_1 metric, it corresponds to a chain-slip between the C and N termini. The metric disconnectivity graph

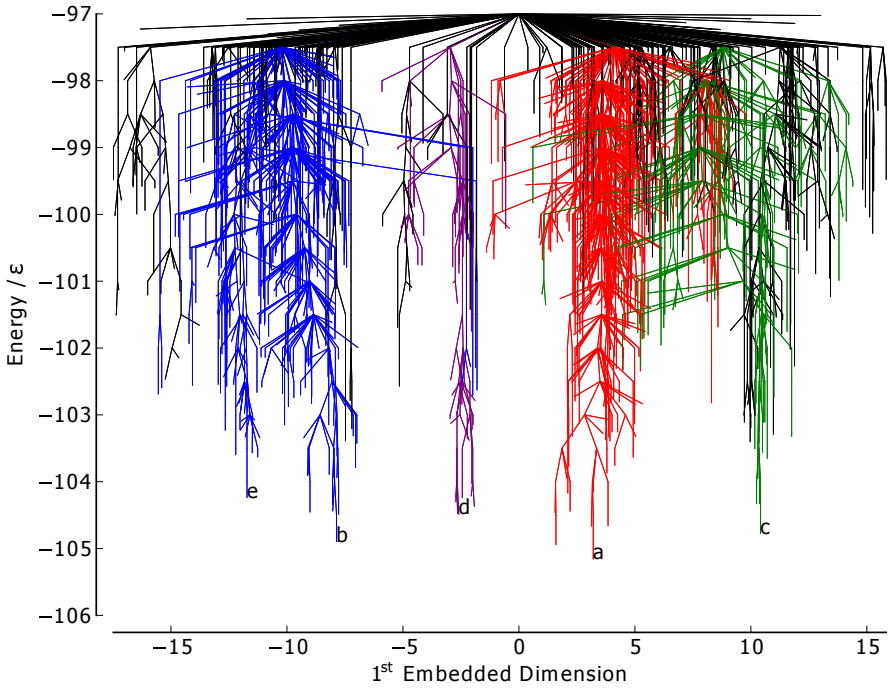


Figure 3.13: Metric disconnectivity graph of BLN-69, $U_t = -97.0\varepsilon$, $N_{sp} = 1611$, with the 1^{st} embedded dimension for $\{\mathbf{r}_\alpha\}_{U_t=-97.0\varepsilon}$ from Isomap analysis used as an order parameter in units of σ. The colour scheme and labels are as used in Figure 3.5.

for the 2^{nd} embedded dimension of the isomap metric (Figure 3.14) is difficult to interpret. That all the coloured funnels overlap suggests that the 2^{nd} embedded dimension corresponds

to some structural variation common to each funnel. The information in Figures 3.9 and 3.10

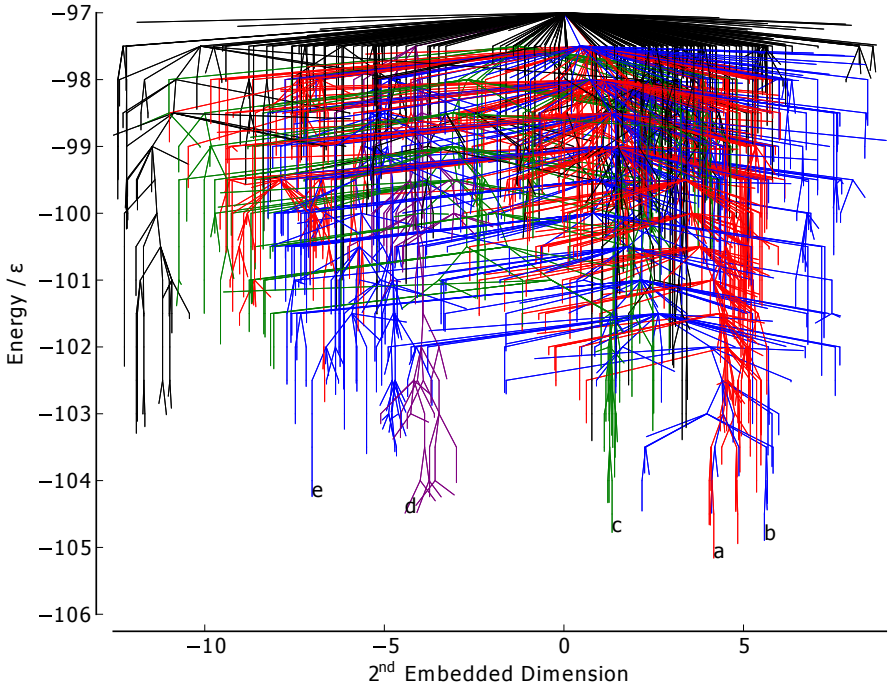


Figure 3.14: Metric disconnectivity graph of BLN-69, $U_t = -97.0\varepsilon$, $N_{sp} = 1611$, with the 2^{nd} embedded dimension for $\{\mathbf{r}_\alpha\}_{U_t=-97.0\varepsilon}$ from Isomap analysis used as an order parameter in units of σ. The colour scheme and labels are as used in Figure 3.5.

is visualised on a single 3D metric disconnectivity graph of $\{\mathbf{r}_\alpha\}_{U_t=-97.0\varepsilon}$ projected onto the plane of maximal variance in Figure 3.15. The plot shows $\{\mathbf{r}_\alpha\}_{U_t=-97.0\varepsilon}$ for BLN-69 plotted against its first two principal components. Clear separation of minima a– e is discernible in this 3D metric disconnectivity graph.

3.3.2 Gō-69

For Gō-69, a database containing 75666 minima and 113101 transition states was used. The first three cartesian principal components for the sets, $\{\mathbf{r}_\alpha\}_{U_t}$, connected to the global minimum below energy U_t, where $-52.0\varepsilon \geq U_t \geq -58.0\varepsilon$ are shown in Table 3.3. As with the cartesian PCs of BLN-69, Q_1 captures significantly more of the variance than Q_2 for all data sets considered. Q_1 for the sub-level set of minima connected to the global minimum below $U_t = -52.0\varepsilon$ and $U_t = -53.0\varepsilon$, have the largest fractional variance, though these large variances are due to a comparatively small number of unstable, high energy minima in

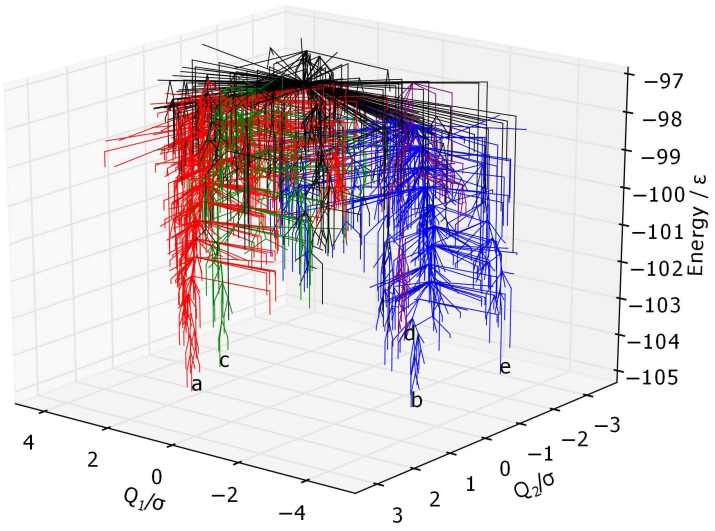


Figure 3.15: 3D metric disconnectivity graph of BLN-69, $U_t = -97.0\varepsilon$, $N_{sp} = 1611$, plotted with the first two principal components of $\{\mathbf{r}_\alpha\}_{U_t=-97.0\varepsilon}$, Q_1 and Q_2, used as order parameters in units of σ. The colour scheme and labels are as used in Figure 3.5.

U_t/ε	N_{sp}	Cartesian PCA			Dihedral PCA		
		S_1^{cart}	S_2^{cart}	S_3^{cart}	S_1^{di}	S_2^{di}	S_3^{di}
−52.0	5529	38.6	14.3	12.6	14.4	10.6	7.3
−53.0	4364	38.0	13.7	12.0	13.7	11.1	7.7
−54.0	3188	24.5	16.0	8.0	13.3	11.6	8.2
−55.0	2386	24.4	16.0	8.1	13.3	11.3	8.7
−56.0	1691	23.7	15.7	7.6	14.0	12.1	9.4
−57.0	1185	21.6	16.3	7.7	13.9	12.7	10.4
−58.0	739	21.5	15.3	7.7	14.9	12.9	11.4

Table 3.3: The variance captured by the first three principal components in cartesian, S_i^{cart}, and dihedral, S_i^{di}, bases of the N_{sp} structures in the sub-level sets of minima below threshold energy U_t for Gō-69.

which one end of the chain has peeled away from the barrel and become unbound. For these systems, Q_1 is no longer representative of the distribution of minima on $U_{\mathbf{r}}$. For this reason, we consider the sub-level set of minima connected to the global minimum below $U_t = -54.0\varepsilon$, for which all the minima have densely packed geometries. This set is represented as a disconnectivity graph in Figure 3.16. The results for the Isomap metric were fairly ambiguous for Gō, with no obvious pattern correlation between the embedded dimensions and and the kinetic or energetic structure of the graph, so they have not been included in this work.

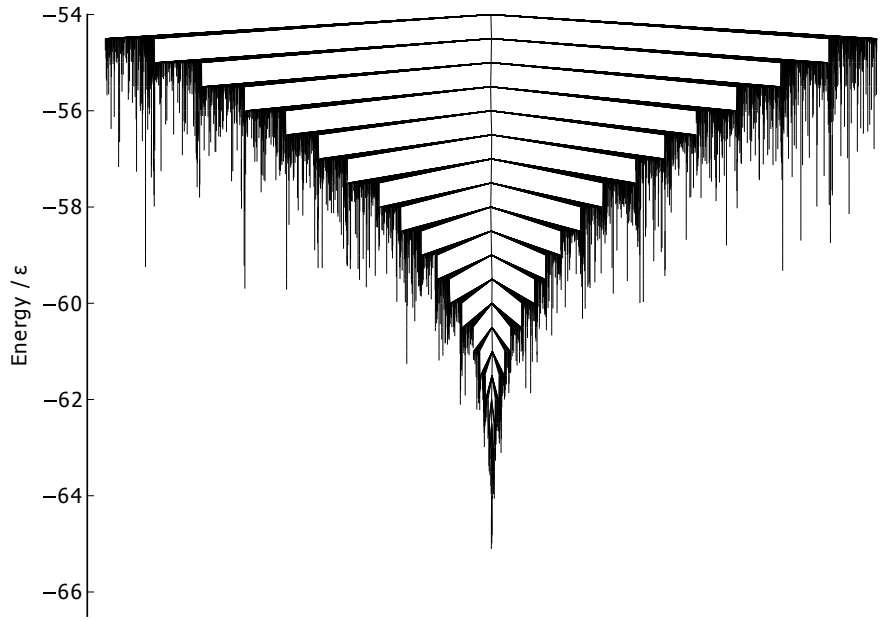


Figure 3.16: Disconnectivity graph of Gō-69, $U_t = -54.0\varepsilon$, $N_{sp} = 3189$.

As there is only a single funnel on the Gō-69 landscape, there are no large kinetic barriers for any of the metrics to differentiate between. The native contact metric (Figure 3.17) is able to partially distinguish between the structures of high and low energy minima. As with BLN-69, minima across the whole energy range examined were able to satisfy nearly full native contacts, including unstable, high energy minima with energetically unfavourable turns in the flexible N bead regions. The converse is not true however, as all low energy minima have a high number of native contacts and low numbers of native contacts are only found for high energy minima.

For the RMSd metric (Figure 3.18) similar behaviour is exhibited as with BLN–69, albeit with a single funnel, with RMSd from the global minimum increasing with increasing energy. The metric disconnectivity graphs in Figure 3.19 use Q_1 and Q_2 as order parameters. In the Q_1 graph the majority of minima are centred about the global minimum, with a smaller number of high energy minima extending to higher values of Q_1. Q_1 and the fraction of native contacts are well correlated, as can be seen in the 3D metric disconnectivity graph (Figure 3.20). The Q_2 metric orders all but a few minima tightly in a rough column about

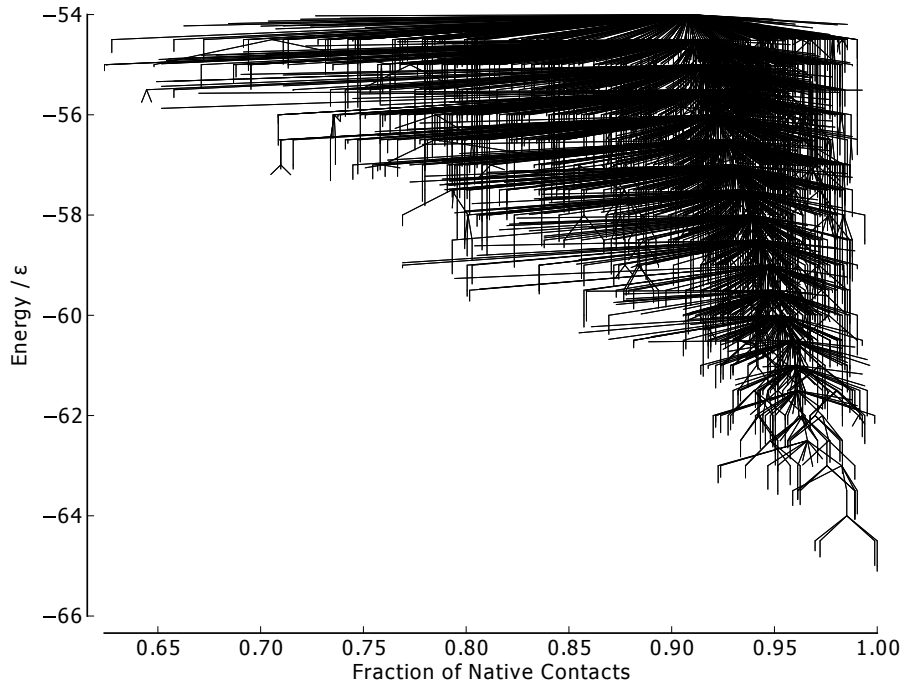


Figure 3.17: Metric disconnectivity graph of Gō-69, $U_t = -54.0\varepsilon$, $N_{sp} = 3189$, with fraction of native contacts used as an order parameter.

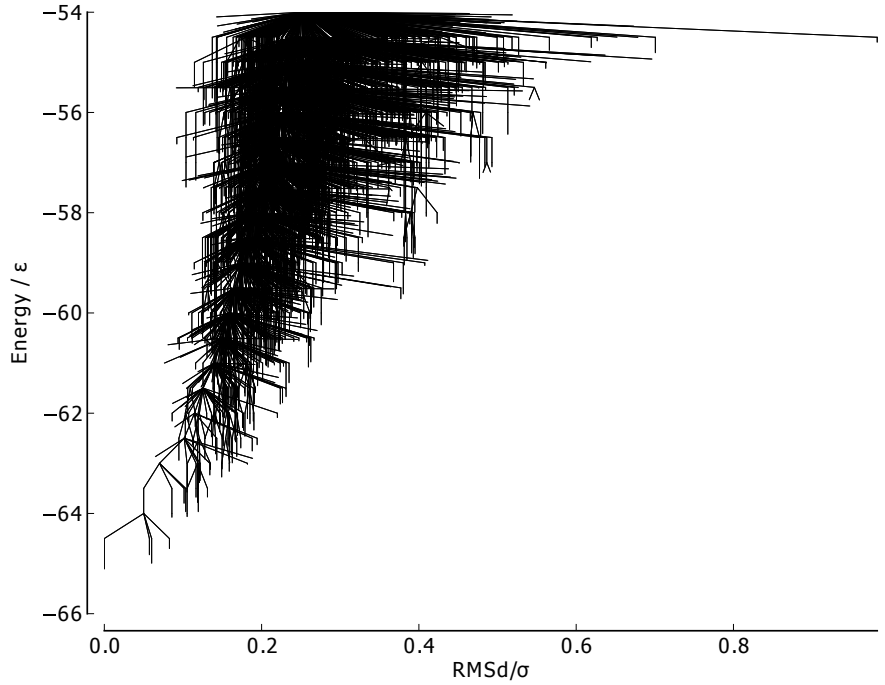


Figure 3.18: Metric disconnectivity graph of Gō-69, $U_t = -54.0\varepsilon$, $N_{sp} = 3189$, with RMSd of each structure from the global minimum used as an order parameter in units of σ

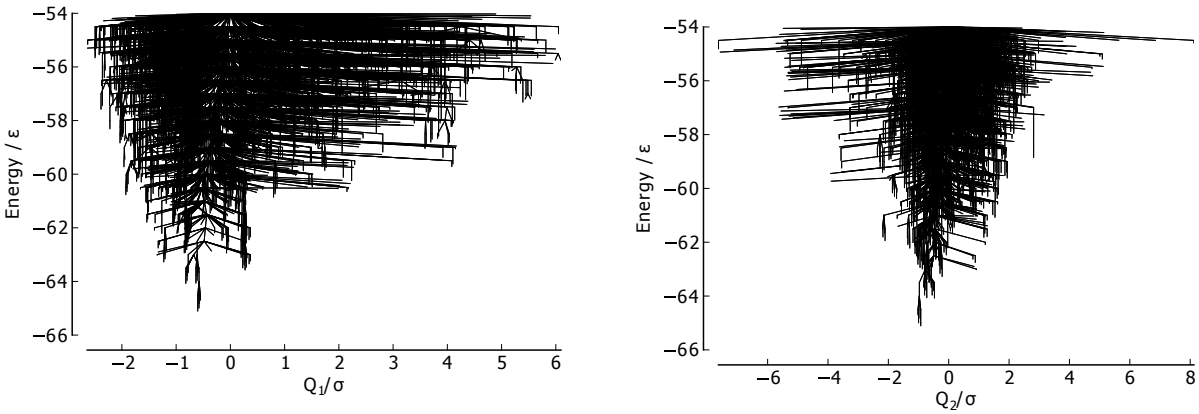


Figure 3.19: Metric disconnectivity graphs of Gō-69, $U_t = -54.0\varepsilon$, $N_{sp} = 3189$, with Q_1 (*left*) and Q_2 (*right*) for $\{\mathbf{r}_\alpha\}_{U_t=-54.0\varepsilon}$, Q_1 and Q_2, used as order parameters in units of σ.

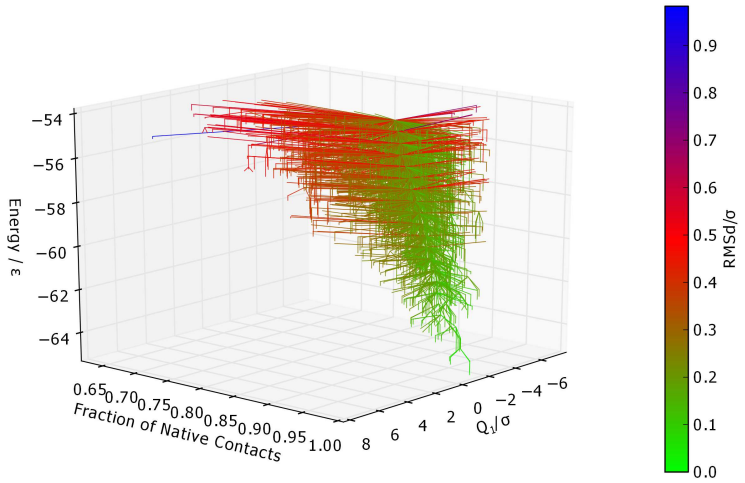
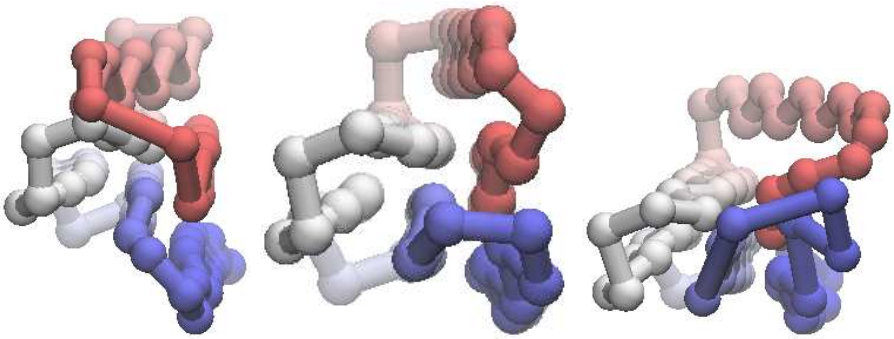


Figure 3.20: Metric disconnectivity graph of Gō-69, $U_t = -54.0\varepsilon$, $N_{sp} = 3189$, with Q_1 for $\{\mathbf{r}_\alpha\}_{U_t=-54.0\varepsilon}$, Q_1, and fraction of native contacts used as order parameters, and coloured according to RMSd of each structure from the global minimum. Q_1 and RMSd are in units of σ.

$Q_2 \approx 0.1$. Those unstable, higher energy minima that are not in that column are structures with a partly unbound C-terminus chain-portion.

The use of colour allows an additional metric to be included on a metric disconnectivity graph. For example, Figure 3.20 shows the Q_1, native contact and RMSd metrics for Gō-69.

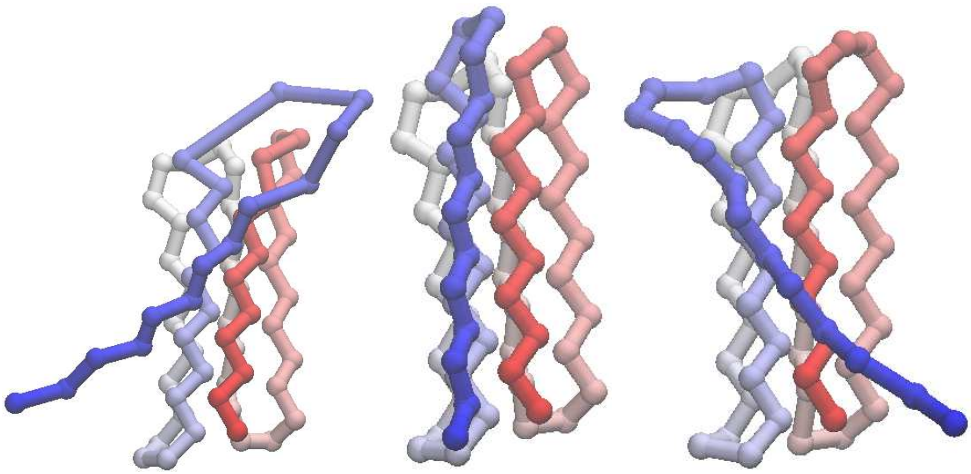
Figure 3.21 shows the progression of Q_1 of $\{\mathbf{r}_\alpha\}_{U_t=-54.0\varepsilon}$ from $\lambda = -2.6\sigma$ to $\lambda = 6.1\sigma$, with a view along the axis of the barrel and corresponds to a sweeping action of the red chain-portion across the face of the white chain-portion. Figure 3.22 shows the progression of Q_2 of $\{\mathbf{r}_\alpha\}_{U_t=-54.0\varepsilon}$ from $\lambda = -7.6\sigma$ to $\lambda = 8.1\sigma$, and corresponds to a "can-can" like



(a) $Q_1 = -2.6\sigma$ (b) $Q_1 = 1.7\sigma$ (c) $Q_1 = 6.1\sigma$

Figure 3.21: Different values of Q_1 for $U_t = -54.0\varepsilon$ projected onto the structure of the global minimum of Gō-69. For the global minimum, $Q_1 = -0.6\sigma$. The beads are coloured from red to blue (N-terminus to C-terminus). An animated version of this projection is available as Supporting Information in [102].

sweeping motion of the free C-terminus end of the chain.



(a) $Q_2 = -7.6\sigma$ (b) $Q_2 = 0.3\sigma$ (c) $Q_2 = 8.1\sigma$

Figure 3.22: Different values of Q_2 for $U_t = -54.0\varepsilon$ projected onto the structure of the global minimum of Gō-69. For the global minimum, $Q_2 = -0.9\sigma$. The beads are coloured from red to blue (N-terminus to C-terminus). An animated version of this projection is available as Supporting Information in [102].

3.4 Conclusions

In this study, we have demonstrated how an appropriate order parameter can elucidate the connection between structures in the energy landscape of BLN-69 and Gō-69, such as funnels, with certain structural motifs of the protein, including chain slips and twists in the turn regions. However, there are still shortcomings to the metrics proposed. Fraction of native contacts and RMSd metrics relied upon having prior knowledge of the system.

PCA provides a means to study systems without resorting to chemical intuition, but still assumes that the point cloud is approximately linear, and cannot be directly implemented for angular co-ordinates. Also, it considers all structures to be of equal importance, regardless of energy, leading to situations such as with Gō-69, where all the variance in structure was provided by a small number of high energy, unstable minima. Isomap allows one to discern low dimensional, non-linear manifolds in the data, and does not make the same assumptions of linearity as PCA. This is clearly a successful strategy, with Isomap distinguishing between the different kinetic structures on the landscape. A useful feature of PCA is the ease with which one can project the principal components back into the original space, making it possible to visualise what these directions correspond to. In principal, it should be possible to do the same with Isomap, projecting the approximate geodesics of the manifold back into the original space, though we have not implemented this in the work presented here.

Other non-linear dimensionality reduction methods exist in the literature, such as sketch-map [103], locally scaled diffusion map [104, 100] and spectral methods [105], which are good candidate metrics for further study. Sketch-map and locally scaled diffusion map have both been designed specifically for the purpose of determining collective coordinates of molecular systems, so they would seem like obvious candidates for use in a metric disconnectivity graph analysis. However, in both cases the required training data are dynamical phase space trajectories, not distributions stationary points in configuration space. If stochastic dynamics are used (such as described by the Fokker-Planck equation) then one does not have to explicitly consider momentum (so the phase space remains in $\mathbb{R}^{3N}$, not $\mathbb{R}^{6N}$) reducing the dimensionality of the problem a little. However, it is not abundantly clear how and if such an approach would work with static data, and should be the subject of future research.

Though the data produced by PyConnect is of a high quality, the data analysis is still fairly qualitative, and further efforts are being taken to quantify the observations, such as using graph-theoretic techniques to analyse and compare tree graphs [106]. Another avenue might be to explore different methods of visualising landscapes away from the disconnectivity approach, such as persistence diagram, which can be used to unambiguously determine the location and frequency with which basins and metabasins appear on a landscape, but at the

expense of explicit connectivity information [107].

Part II

Hydrated Ions

Chapter 4

Motivation, Models and Methodology

This chapter introduces micro-hydrated ions as systems of interest for both experiment and computation. The micro-hydrated ions is a single ion coordinated by a finite number of water molecules, and takes the general form $X(H_2O)_n$. Micro-hydrated ions are attractive systems to study as they can be used to probe the manner and extent to which an ion can order its neighbouring solvent molecules, which can be used to rationalise various ion-specific effects in physical chemistry.

4.1 Hydrated Anion Clusters

In this section a brief discussion of two topics is presented. Firstly, a review of current experimental and simulation work concerning hydrated anions is given (cations are not considered in this thesis), and secondly the *Hofmeister series* is introduced, the origins of which are a consequence of ion-induced solvent ordering.

Mass spectrometrists first began to produce gas phase hydrated anions in experiment in the late 1960s. By using an alpha particle ionisation mass spectrometer, Kerbale *et al.* [108] were able to produce singly charged anions such as F^- and Cl^-. When these species were introduced to water vapour, the neutral water molecules were observed to aggregate about

the ions, hydrating them. It soon became clear that micro-hydrated ions comprising a single solute anion and a few solvent water molecules could provide insight into the solvation properties of ions in the bulk, and interest blossomed [109]. Numerous techniques and methods for producing size-selected hydrated anion clusters are now available to the modern researcher [110], including electrospray ionisation [111], sputtering of frozen solutions [112], laser vaporisation [113] and ion injection from thermionic emitters [114].

Not withstanding their inherent interest, hydrated halide ions are a thoroughly explored and well characterised set of singly charged anionic systems due to their role as an ideal model system with which to construct and evaluate water-ion interaction potentials. Argon predissociation spectroscopy and infrared photodissociation spectroscopy of gas-phase hydrated halide ions indicate that the ions prefer to sit on the surface, with the exception of F^-, which disrupts water-water interactions and sits centrally [115, 116, 114]. Similar ordering behaviour is also replicated in ab initio calculations [117, 118] and molecular dynamics simulations [119, 120]. The strong ordering effect of the F^- ion on surrounding water molecules manifests itself in bulk solution also. The residence time of water molecules in the first solvation shell is approximately 1.5 times greater for the fluoride ion then for the other halides [121]. This in turn results in an anomalously small diffusion constant for the fluoride ion in dilute aqueous solution when compared to the other halides, as it is obliged to drag itself *and* a tightly bound inner solvation layer through the solution [121, 122].

Multiply charged anions (MCA) pose a different set of challenges to singly charged anions, as they are fragile and prone to decay due to the strong, repulsive interactions between the excess charges on the ion. When present in condensed phases, MCAs are stabilised by dielectric screening which tempers the coulombic interactions. However, when in the gas phase such screening effects are absent, and so MCAs are susceptible to decay by either electron detachment or ionic fragmentation [123].

The stability of an MCA can be characterised in three ways; vertical electronic stability, adiabatic electronic stability and thermodynamic stability. An MCA, X^{n-}, is vertically electronically stable if, when at its optimal geometry, it is lower in energy than $X^{(n+1)-}$ at the same geometry. An adiabatically electronically stable MCA is one for which the X^{n-} op-

timal geometry is more stable than the $X^{(n+1)-}$ species at its own optimal geometry. For an MCA to be thermodynamically stable, the $X^{(n)-}$ optimal geometry must be more stable than all other possible fragments (thermodynamic stability implies both vertical and adiabatic electronic stability). Most MCAs are not thermodynamically stable, though they are often electronically stable and thus support a bound state. In spite of their lack of thermodynamic stability, such systems can be long-lived due to another notable feature of MCAs; the *repulsive coulomb barrier* (RCB). RCBs are characterised by a repulsive interaction at medium-and-long-distances between like charges (i.e. between two ion fragments or between the ion and free electron) and a short-range attractive interaction due to the presence of a bound state. There are thus two minima (corresponding to the bound state and the limit of infinite separation of the charge carrying bodies, respectively) which are separated by a barrier (the RCB). Providing that the bound state minimum is sufficiently deep to support

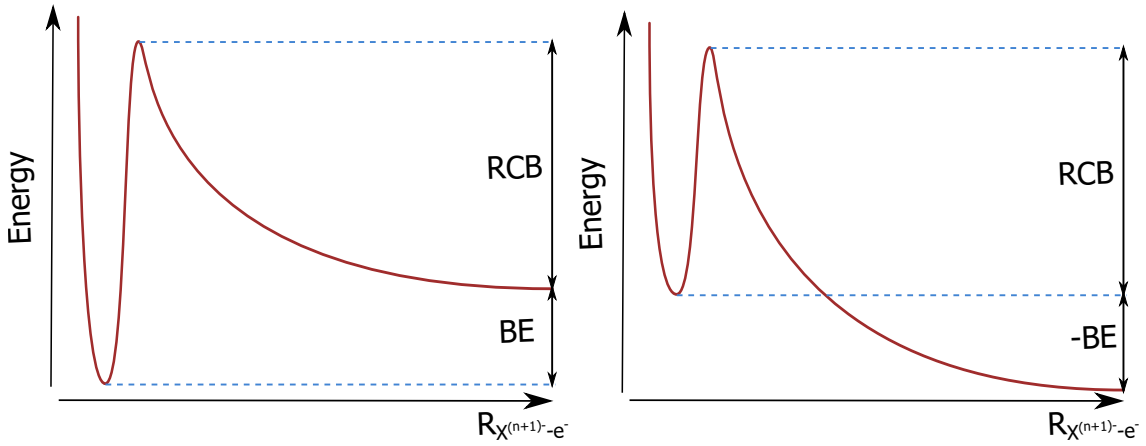


Figure 4.1: Schematic potential energy curves representing a bound state with electron binding energy BE and the RCB for an MCA. The figures on the left and right represent a thermodynamically stable MCA and a metastable MCA, respectively.

the zero point vibrational states and the RCB is sufficiently wide and high to inhibit fragmentation or autodetachment, the MCA can exist as a metastable state (see Figure 4.1). The RCB with respect to electron detachment has been measured experimentally by the group of Lai-Sheng Wang, who studied the photoelectron spectra of the citrate dianion [124] and determined the height of the RCB to be between 1.9 and 2.5 eV. It is also possible to directly detect metastable MCAs from their photoelectron spectra, where the metastability manifests itself as a negative electron binding energy (the experimental signature is that the

ejected electrons are observed to have a *greater* kinetic energy than the incoming photons). MCAs for which negative electron binding energies have been observed include the copper phthalocyanine tetrasulfonate tetraanion ($[CuPc(SO_3)_4]^{4-}$) [125, 126] and the $PtCl_4^{2-}$ and $PtBr_4^{2-}$ dianions [127]. The RCB with respect to ion fragmentation has also been detected experimentally for the $IrBr_6^{2-}$ dianion, where photodissociation results in Br^- and $IrBr_5^-$ fragments [128]. The RCBs for ion fragmentation and electron detachment were measured to be approximately 1.6 and 2.2 eV, respectively, indicating that for $IrBr_6^{2-}$ ion fragmentation is the lowest energy unimolecular decay pathway [123].

MCAs which are either electronically or thermodynamically unstable in isolation (such as SO_4^{2-}, $C_2O_4^{2-}$ and $Fe(CN_6)^{3-}$) can be stabilised by solvation or counter-ion complexation. Unlike the singly charged anions, unstable gas phase MCAs cannot be prepared as bare ions and then subsequently hydrated. Rather, gas phase hydrated clusters or ion-pair complexes must be produced in a single step using a technique such as electrospray ionisation [111]. An example of a counter-ion complex which is stable in the gas phase is the SO_4^{2-} dianion, which is stabilised with respect to autodetachment by coordinating it with an alkali metal ion such as Na^+ or K^+ [127]. A more widely studied case is the stabilisation of an MCA with respect to autodetachment via solvation. Determining the minimum number of water molecules required to stabilise a given gas phase MCA is an active topic of interest amongst spectroscopists. Photoelectron spectroscopy of size selected sulfate and oxalate dianions reveals that a minimum of three water molecules are required to electronically stabilise each of the ions [126, 129]. Infrared photodissociation spectroscopy of the hydrated ferricyanide trianion indicates that a minimum of seven water molecules is necessary to stabilise it [130]. As the number of solvent molecules increases, focus shifts from the electronic structure of the ion (and the stabilisation of the ion by the solvent) to the geometric structure of the cluster. Questions of interest typically regard the location of the ion within the cluster (i.e. does it sit centrally or on the surface), the network of hydrogen bonds within the cluster and the appearance and size of hydration shells. Examples of such include infrared multiple photon dissociation studies of the bicarbonate and bisulfate anions up to 10 and 12 water molecules, respectively [131, 132], and ab initio structure prediction of hydrated hydride and

hydroxide clusters up to six water molecules [133, 134].

Due to its familiarity as a common inorganic MCA and the role it plays in atmospheric chemistry [135] and a number of industrial processes, the hydrated sulfate dianion has been the subject of a great deal of research interest. A number of experimental studies have investigated the solvent structure of the hydrated sulfate anion using photoelectron spectroscopy and photodissociation spectroscopy of size selected clusters, and have been able to determine that; the sulfate ion sits at the centre of the cluster [126, 129], it is capable of patterning the local water structure into trimeric water rings [136], and the first solvation shell closes at around twelve water molecules [137]. Simulation has also been active in determining the solvent structure about the sulfate, with DFT-based Born-Oppenheimer and Carr-Parrinello molecular dynamics used to investigate structural properties and anharmonic corrections to the infrared spectra, and to investigate sulfate polarisability, respectively [138, 139]. Ab initio structure determination has also been conducted up to seven water molecules [140]. The extent and manner to which the sulfate ion is able influence the surrounding solvent structure is of particular interest, as it can be used to rationalise the position of the ion along the Hofmeister series (see Subsection 4.1.1). Infrared photodissociation spectroscopy of size selected hydrated sulfate clusters in the range 36 - 80 water molecules has been used to probe the long range water ordering due to the sulfate ion [141]. The spectra reveal that dangling OH bonds, a feature common in pure water clusters of all sizes and at the bulk-vapour interface, are suppressed in hydrated sulfate clusters. For clusters containing less than 47 water molecules the experimental signature of a dangling OH bond (a peak in the intensity of the infrared photodissociation spectra at 3710 cm^{-1}) is not observed. For 47 water molecules and larger the signal is present, and grows more intense with increasing cluster size. This suggests that the sulfate ion is capable of influencing its surrounding water structure beyond the first solvation shell (approximately 12 water molecules) and well into the second or third shell. This observation is corroborated by our own simulation data of hydrated sulfate clusters [142], the results of which are discussed in depth in Chapter 5. The hydrated ferricyanide ion also exhibits a similar size dependent suppression of dangling OH bonds, though with the dangling OH bond signal first appearing at approximately 80 water

molecules [130]. Simulation results for the hydrated ferricyanide ion provisionally agree with this data, and are reported and discussed in Chapter 6.

4.1.1 Hofmeister Series

The Hofmeister series ranks anions and cations according to their relative effectiveness at influencing a number of interesting phenomena in physical chemistry [143, 144]. Within the series, ions can be characterised as being either *kosmotropic* (structure making) or *chaotropic* (structure breaking). Kosmotropes increase the surface tension of liquids [145] and the stability of proteins [146], and decrease the solubility of hydrophobic particles [147] and the denaturation of proteins [148], (to name but a few properties), whilst chaotropes behave in the opposite manner [149].

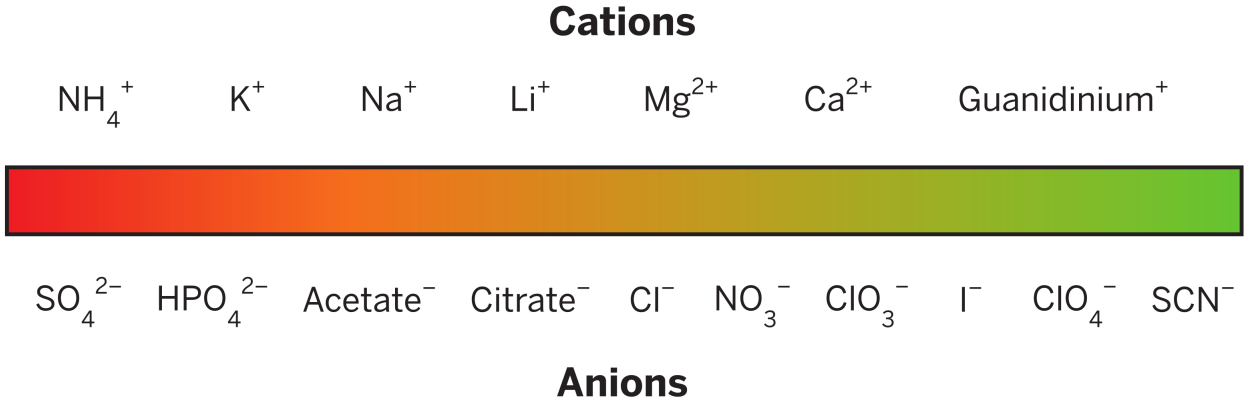


Figure 4.2: The Hofmeister series. The scale runs from kosmotropic (red,left) to chaotropic (green,right). Adapted from [149].

Despite the ubiquity of Hofmeister effects the underlying chemical origins remain unclear, with theory and experiment seemingly providing evidence for two competing mechanisms; long range restructuring of the surrounding solvent molecules by the ions and direct interaction with the electric field of the ion [150, 151, 152, 145].

Historically, it was believed that the extent to which an ion restructured its surrounding solvent was the determining factor for its position on the Hofmeister series. By referencing the number of solvation shells which were believed to be altered by an ion, Hofmeister ions were characterised as being either polar kosmotropes, chaotrope or non-polar kosmotropes [153]. However, a number of experiments began to cast doubt on this interpretation, pointing to al-

ternative explanations for the observed ion specific effects. By using femtosecond two-colour pump-probe spectroscopy, Bakker and co-workers were able to investigate water reorientation dynamics in salt solutions [154]. These experiments showed that although the orientational correlation times for waters in the first solvation shell of a number of Hofmeister ions were much longer than for the bulk, the effect did not extend to successive layers. Even at high concentrations, neither strongly kosmotropic ions (sulfate) nor strongly chaotropic ions (perchlorate) exhibited depressed correlation times in the second solvation layer or beyond. Put otherwise, regardless of the supposed structuring prowess of a given ion, solvent structure was seemingly unperturbed beyond the first layer of water molecules. Other experimental data which favour an alternate explanation for the Hofmeister series include pressure perturbation calorimetry of Hofmeister salt solutions [155] and vibrational sum frequency spectroscopy of octodecylamine monolayers sat at the surface of a salt solution [156], neither of which argues in favour of the solvation structuring hypothesis.

On the other hand, evidence exists for Hofmeister ions having a perturbative influence on water structure in the second solvation shell and beyond. Sum-frequency generation spectroscopy, a second-order optical technique sensitive to the first few water molecules at the liquid-vapour interface, can be used to investigate the effects of salt on the structure of water at aqueous interfaces (phenomena which appear to follow a Hofmeister-like series). Data from such experiments suggest that bromide and iodide ions are highly surface active whereas other ions such as chloride and sulfate are less so, a result which mimics their Hofmeister ordering [157, 158].

Gas phase hydrated ion clusters are ideal systems for investigating ion-induced water structure, as it is possible to probe the influence of a single ion, and the effect of counter ions can be eliminated. Infrared photodissociation spectroscopy can be used to determine both the internal and external hydrogen bond network in a hydrated cluster, and in a more direct manner than non-linear spectroscopic methods such as sum-frequency generation spectroscopy. A large number of hydrated ion species of varying sizes and solvent numbers have been investigated using this method [159]. As discussed in Section 4.1, of particular interest to understanding long range solvent ordering is the size at which dangling OH bonds begin

to appear at the surface of the cluster (or put another way, for how long is an ion able to suppress their appearance). For the sulfate and ferricyanide ions, dangling OH bonds do not begin to appear until above 43 and 70 water molecules, respectively, well beyond the first (and probably second) solvation shells [141, 130]. As pure water clusters exhibit dangling OH bonds at all sizes [160], this result indicates that the ions are having a significant effect on the both the local and non-local water structure.

4.2 Rigid-body Modelling

Due to the relatively large system sizes under consideration, the hydrated ion clusters were modelled using an empirical potential with bond lengths and bond angles held rigid. The binding energy, U, of a cluster containing N atoms was computed as a sum of pairwise Coulombic and Lennard-Jones terms,

$$U = \sum_{i=1}^{N} \sum_{j=i+1}^{N} \left\{ \frac{q_i q_j}{r_{ij}} + 4\varepsilon \left[\left(\frac{\sigma}{r_{ij}} \right)^{12} - \left(\frac{\sigma}{r_{ij}} \right)^{6} \right] \right\}, \tag{4.1}$$

where i and j are the atom indices, q_i is the partial charge on atom i, r_{ij} is the distance between non-bonded atoms i and j, and σ and ε are Lennard-Jones parameters. Lennard-Jones parameters for interactions between unlike atom types were calculated using the Lorentz-Berthelot mixing rules.

Angle-axis variables were used to describe the rigid-body rotational coordinates, which have been shown to be as efficient as Euler angles, while avoiding the problems in geometry optimisation that arise from singularities in that coordinate system [161, 162]. The angle-axis representation of a rigid body is based on Euler's theorem [163], which states that the general displacement of a rigid body with one point fixed is a rotation about some axis. The angle-axis representation employs a single (unnormalised) vector $\mathbf{p} = [p_1, p_2, p_3] = \theta\hat{\mathbf{p}}$, the direction ($\hat{\mathbf{p}}$) and magnitude ($\theta$) of which define the axis of rotation and the magnitude of rotation about that axis respectively. Therefore, the configuration space of a system of N (non-linear) rigid bodies requires $6N - 6$ degrees of freedom, thus $\mathbf{r} = [\mathbf{x}, \mathbf{p}]$, where $\mathbf{x} = [x_1, \ldots, x_{3N-3}]$ and $\mathbf{p} = [p_1, \ldots, p_{3N-3}]$ are the position vector of the centres of mass

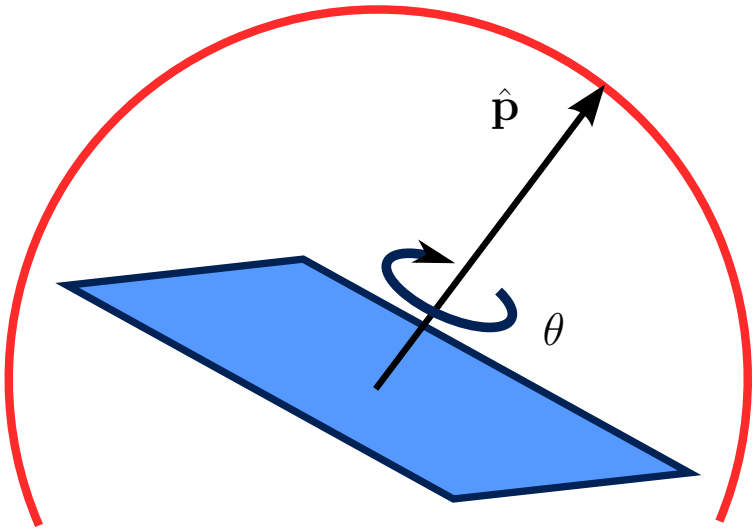


Figure 4.3: Schematic of the angle-axis representation of a rigid-body rotation. The blue rectangle represents a rigid body which is being rotate about axis $\hat{\mathbf{p}}$ by a magnitude θ. Because $\hat{\mathbf{p}}$ is a unit vector, it is constrained to lie on the surface of the unit sphere.

and the rotation vector for of each body, respectively.

4.2.1 Rigid-body Modelling of Water Molecules

The water molecules are represented by the four-site rigid-body TIP4P water potential [164], which includes the two hydrogen atoms, the oxygen atom and a lone pair site. The O-H bond distance and H-O-H bond angle are 0.9572 Å and 104.52°, respectively, and the lone pair site lies 0.15 Å along the H-O-H bond angle bisector. The potential parameters are listed in Table 4.1.

Atom	q_i / e	σ / Å	ε / kcal mol^{-1}
Hydrogen	+0.52	0.0	0.0
Oxygen	0.0	3.15	0.648
Lone Pair	-1.04	0.0	0.0

Table 4.1: TIP4P potential parameters for H_2O [164]

TIP4P water clusters have been studied extensively [165, 166, 167, 168, 169, 170, 171, 160, 172, 173], with good candidate global minima determined for up to 36 water molecules [169], and reasonable agreement with both *ab initio* calculations and experiment up to 12 molecules [171]. TIP4P water is also able to qualitatively describe the water phase diagram [174, 175], though modifications to the model exist which can give better quantitative agreement [176].

4.2.2 Rigid-body Modelling of the Sulfate Ion

The sulfate dianion is modelled as a tetrahedral molecule with rigid S-O bonds of length 1.49 Å and O-S-O bond angles of 109.5°. The sulfate potential that we have used was derived by

Atom	q_i / e	σ / Å	ε / kcal mol^{-1}
Sulfur	+2.4	3.55	0.25
Oxygen	-1.1	3.15	0.25

Table 4.2: Potential parameters for SO_4^{2-} [177]

fitting Møller–Plesset MP4SDTQ level calculations by McCammon *et al.* and has been shown to reproduce experimental solution data [177]. Previous calculations of the micro-solvation of the sulfate ion at the MP2 level also agree with the MP4 results [178]. McCammon *et al.* suggested two sulfate potentials, with the same Lennard-Jones parameters but different partial charges on the sulfur and oxygen atoms. In order to choose between these parameter sets, we performed short basin-hopping runs for $SO_4^{2-}(H_2O)_n$ in the size range $3 \leq n \leq 7$ for both potentials. The 40 lowest-energy isomers for each n were then reoptimised using density functional theory (DFT) level with the B3LYP exchange-correlation functional and a 6-311++G** basis set, as implemented within the NWChem package [23]. The choice of exchange-correlation functional and basis set is discussed in Section 5.5. Little difference in isomer ordering was found when comparing the two potentials, so we chose to use the first of the two parameter sets suggested by McCammon *et al.* (Table 4.2) because it had a slightly better Spearman rank correlation with the DFT calculations.

4.2.3 Rigid-body Modelling of the Perchlorate Ion

The perchlorate anion is modelled as a tetrahedral molecule with rigid Cl-O bonds of length 1.44 Å (taken from scattering data) and O-Cl-O bond angles of 109.5°. The perchlorate

Atom	q_i / e	σ / Å	ε / kcal mol^{-1}
Chlorine	+0.44	4.9	0.0401572
Oxygen	-0.35	3.1	0.075723

Table 4.3: Potential parameters for ClO_4^- [179]

potential that we have used was derived from an LCAO MO SCF calculation by Heinje

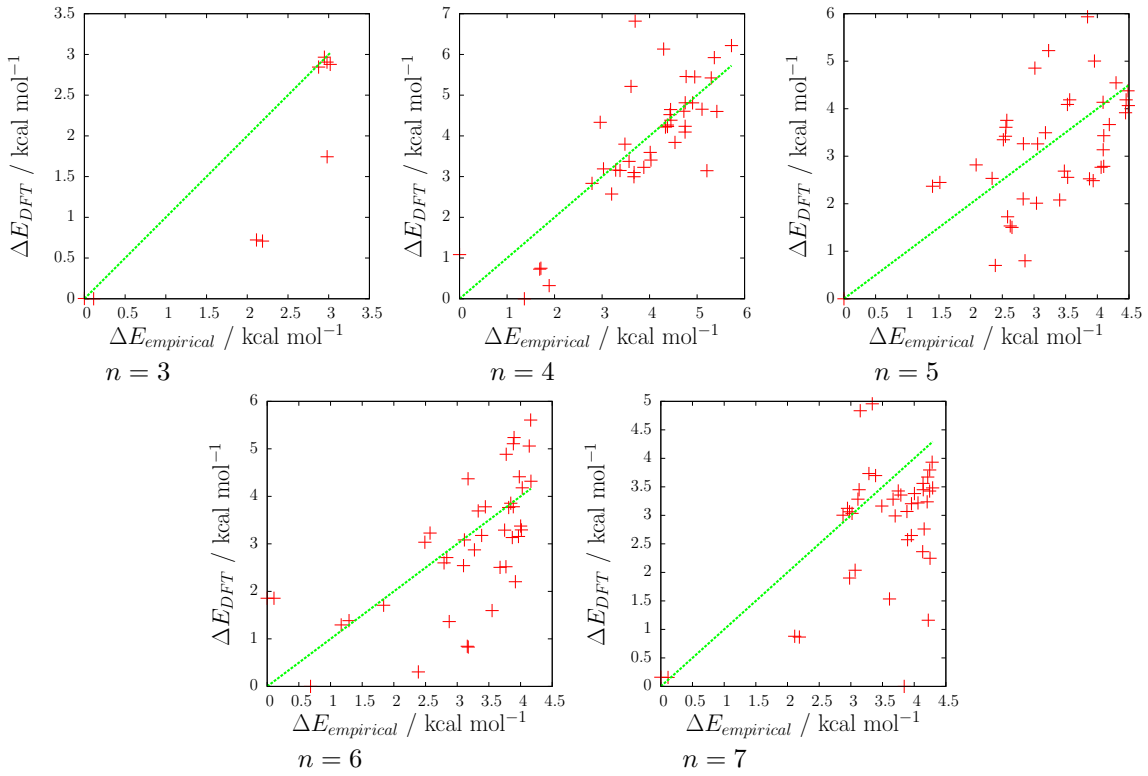


Figure 4.4: Correlation plot of reoptimised DFT minima computed at the B3LYP/6-311++G** level of theory against the original rigid-body minima, for $SO_4^{2-}(H_2O)_n$, with $3 \leq n \leq 7$.

et al. [179], who used it to study a 1:1 aqueous salt solution of $NaClO_4$ using molecular dynamics simulations. The vibrational spectra of both the bare and hydrated perchlorate anion have been investigated at the ab initio level using a stepping stone approach, where the cluster geometries are reoptimised at successively higher levels of theory, spanning from HF/STO-3G to MP2/6-31+G* [180].

4.2.4 Rigid-body Modelling of the Thiocyanate Ion

The thiocyanate anion is modelled using the 5-site planar model of Wipf *et al* [181], which includes sulfur, carbon and nitrogen atoms (arranged linearly) and two lone pair pseudo-atoms orthogonal to the SCN axis, either side of the sulfur atom. The S-C, C-N and LP-S distances are 1.59 Å, 1.25 Å and 0.65 Å, respectively. The model charges were derived from a gas phase SCN^- anion optimised at the Hartree-Fock level with a 6-31+G* basis set, with the sulfur charge split equally between the two lone pair sites. For the Lennard-Jones parameters, the OPLS acetonitrile force field [182] was used for the nitrogen and carbon

and the AMBER94 force field [183] was used for the sulfur atom and lone pair site. The charge and Lennard-Jones parameters were further modified using a trial-and-error method to improve the structures and relative stabilities of $La(NCS)_6^{3-}$ and $La(SCN)_6^{3-}$ complexes in the gas phase.

Atom	q_i / e	σ / Å	ε / kcal mol^{-1}
Sulfur	0.00	4.1426	0.200
Carbon	0.40	3.340	0.086
Nitrogen	-0.70	3.271	0.170
Lone Pair	-0.35	2.575	0.155

Table 4.4: Potential parameters for SCN^- [181]

4.2.5 Rigid-body Modelling of the Ferricyanide and Ferrocyanide Ions

The ferricyanide ($Fe(CN)_6^{3-}$) and ferrocyanide ($Fe(CN)_6^{4-}$) anions are both modelled using a 13-site octahedral model developed by Wang *et al* [184]. The models were used to simulate the structures and dynamics of the ions in water and heavy water, and were able to reproduce experimental features seen in 2D infra-red spectroscopy. For the ferricyanide ion, the Fe-C and C-N bond lengths are 1.997 Å and 1.176 Å, respectively, and for the ferrocyanide ion the Fe-C and C-N bond lengths are 2.035 Å and 1.186 Å. For each ion, equilibrium bond lengths, angles and dihedrals were derived from QM optimised geometries using the JOYCE [185, 186] parametrisation procedure. Point charge distributions were obtained using the CHelpG procedure at the B3LYP/6-31+G* level. For the Lennard-Jones parameters, the Fe was modelled with a modified UFF potential [187, 188], and the C and N atoms were modelled with the OPLS-AA forcefield [189]. Both ions were assigned the same Lennard-Jones parameters.

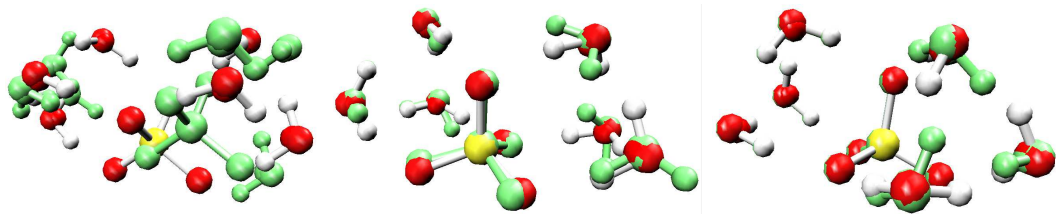
Atom	q_i / e	σ / Å	ε / kcal mol^{-1}
Iron	-0.834 / -1.150	4.9	0.0401572
Carbon	0.403 / 0.431	3.1	0.075723
Nitrogen	-0.763 / -0.906	3.1	0.075723

Table 4.5: Potential parameters for $Fe(CN)_6^{3-}$ / $Fe(CN)_6^{4-}$ [184]

4.3 Global Optimisation of Hydrated Ions

For each cluster in the size range $3 \leq n \leq 50$, eight independent searches were conducted starting from different random geometries. Searching was continued until either; each of the eight independent basin-hopping simulations had found the same common lowest energy structure, or each had taken ≈1.5 million basin-hopping steps. For the larger cluster sizes this was not always possible, and those structures which did not meet this requirement are indicated in the results chapters.

As discussed in Chapter 2, choosing a move class from which to draw random perturbations to the geometry and choosing a temperature to use for the metropolis acceptance criterion are critical to the efficiency of the basin-hopping simulation. For the hydrated ion global minima presented in this thesis, three distinct move classes were employed; rigid-body translations of the centre-of- mass of the ion and water molecules, rigid-body rotations about the oxygen atom (for the water molecules) or central atom (for the ion) and *cycle inversion* moves.



(a) Rigid-body translation (b) Rigid-body rotation (c) Cycle inversion

Figure 4.5: The three move classes from which geometry perturbations are drawn during a basin-hopping simulation, as demonstrated for $SO_4^{2-}(H_2O)_6$. In each of the figures, the true-coloured species represent the starting minima, $\mathbf{r}_\alpha$, and the green structures are the result of adding a random perturbation, $\mathbf{r}_\alpha + \delta\mathbf{r}$.

Rigid-body translations and rotations are the most fundamental way to perturb the geometry of a rigid-body system. During a rigid-body translation, the centre of mass of each rigid body is perturbed by a random vector, $\delta\mathbf{x} = x_{step}\delta\hat{\mathbf{x}}$, where $\delta\hat{\mathbf{x}}$ is drawn uniformly from the surface of a sphere, i.e.;

$$\delta\hat{\mathbf{x}} = [\delta\hat{x}_1, \delta\hat{x}_2, \delta\hat{x}_3] = [\sqrt{1-z^2}\cos\phi, \sqrt{1-z^2}\sin\phi, z] \tag{4.2}$$

where $z \in [-1, 1]$ and $\phi \in [0, 2\pi]$ are uniformly distributed random variables. For a rigid-body rotation, the centre of mass of each rigid body is perturbed by first choosing a random

axis of rotation in the manner described in Equation 4.2, and then randomly choosing a magnitude of rotation, θ. Simply drawing θ from a uniform distribution will not lead to uniform random rotation displacements, and one must instead choose θ according to;

$$p(\theta) \propto \frac{1}{\pi} \sin^2\left(\frac{\theta}{2}\right) \tag{4.3}$$

for $\theta \in [0, \theta_{step}]$. The parameters x_{step} and θ_{step} control the magnitude of the translational and rotational perturbations, and are key to the efficacy of a move class. For the results presented in this thesis, x_{step} and θ_{step} were allowed to change throughout the course of the search according to an adaptive stepsize-and-temperature basin-hopping algorithm described in Subsection 4.3.1.

In water-rich systems there are a large number of structural isomers which share the same oxygen skeleton (the positions of the oxygen atoms) and only differ from one another by the placement of the hydrogen atoms. The strong directional interactions which arise from hydrogen bonding between water molecules means that isomers with the same oxygen skeleton but different hydrogen bonding networks can differ greatly in energy. Therefore, one should expect moves classes which explore the hydrogen bonding topology of a given oxygen skeleton to be useful. Cycle inversion moves (equivalent to Takeuchi's closed chain perturbations [166]) attempt to alter the hydrogen bonding network in the cluster whilst preserving the oxygen skeleton. A cycle in a directed network is a closed loop of edges with the direction of each edge pointing the same way around the loop [190]. Such a loop can be inverted by pointing edges in the opposite direction. The cycle inversion move class searches for simple cycles (those in which no node appears twice) in the directed network of water-water hydrogen bonds using the `simple_cycles` method in `networkx` [191], a Python library for the creation, manipulation and study of networks. In a cycle inversion move, a cycle is chosen at random from the hydrogen bond network and inverted. If no cycles are present, a rigid-body translation or rotation is performed instead. A water molecule i is considered to donate a hydrogen-bond to water molecule j if the oxygen-oxygen distance is less than 3.50 Å and the OH bond axis of molecule i is within 30° of the displacement vector between the oxygen

atoms in i and j. Some structures with positively charged atoms close to the TIP4P lone pair underwent cold fusion. These structures were discarded from the basin-hopping search.

Global optimisation studies of pure water clusters have suggested that the use of *block moves*, where one type of move class is used exclusively for a set number of steps before switching to another, is an effective strategy for finding low energy structures [160]. In this study, we used blocks of 100 steps for each type of move.

4.3.1 Adaptive Stepsize-and-Temperature Algorithm

In the simplest implementation of a basin-hopping algorithm, one would choose values for the temperature and stepsize, and hold them constant throughout the course of the simulation. This strategy can lead to inefficient searching, as there is no guarantee that the magnitude of the random geometry perturbations and the temperature will be compatible with the landscape topography. One can improve the search efficiency by "tuning" the basin-hopping parameters (i.e. performing a parameter sweep of the stepsize and temperature, and selecting the most successful combination), but this can be expensive, and assumes that a single set of parameters will be appropriate for the whole landscape.

The adaptive stepsize-and-temperature algorithm used to generate the results in the following chapters circumvents this problem by allowing the temperature and stepsize to vary during the search. The two core tenets of the algorithm are that the basin-hopping parameters should be adapted in order to maintain a constant ratio of accepted-to-rejected basin-hopping steps, R, and that the magnitude of $\delta\mathbf{r}$ and T can be treated independently. The geometry perturbations and the metropolis temperature can be decoupled due to $\delta\mathbf{r}$ being solely responsible for moving from one basin to another, whilst T alone controls how likely such a transition will be accepted. Formally, we define the step ratio

$$R_{step} = \mathrm{p}\left(\mathbf{r}_\alpha \neq \mathbf{r}_\beta\right) \tag{4.4}$$

to be the probability that a basin-hopping step moves from one minimum to another. The geometries $\mathbf{r}_\alpha$ and $\mathbf{r}_\beta$ were considered equivalent if $|U(\mathbf{r}_\alpha) - U(\mathbf{r}_\beta)| \leq 10^{-3}$ kcal mol^{-1}.

The temperature ratio,

$$R_{temp} = \mathrm{p}\left(\alpha \rightarrow \beta | \mathbf{r}_\alpha \neq \mathbf{r}_\beta\right) \tag{4.5}$$

is the probability that the move from $\mathbf{r}_\alpha$ to $\mathbf{r}_\beta$ is accepted. The total acceptance ratio is then;

$$R = 1 - R_{step}\left(1 - R_{temp}\right) \tag{4.6}$$

which includes the probability of remaining in the same minimum. In the results presented, the target values of R_{step} and R_{temp} were 0.8 and 0.3 respectively, yielding an overall acceptance ratio of about 50%. For a given move class, R_{step} and R_{temp} were dynamically adjusted towards the target values after a period of 100 steps (the length of a block of moves). Furthermore, each of the move classes were allowed to vary their stepsize and temperature independently. The consequence of this decoupling of move class parameters means that each type of move is able to search the landscape with a unique temperature and magnitude of perturbation (hopefully) tailored to the landscape topography experienced whilst exploring according to this class.

4.3.2 Search Strategy Performance

The adaptive stepsize-and-temperature algorithm described in Section 4.3.1 was benchmarked against two different implementations of the basin-hopping algorithm. The first method utilised only rigid-body rotations and translations performed simultaneously each step (i.e. no block moves and no cycle inversion moves), with a constant temperature and stepsize throughout the simulation. The second method utilised block moves of rigid-body translations, rotations and cycle inversions, again with the stepsize and temperature constant throughout. In order to (approximately) gauge the efficacy of each method as a global optimisation strategy, the microhydrated sulfate ion was used as a test system. To assess the performance of each method, eight independent searches were conducted from different random $SO_4^{2-}(H_2O)_n$ starting geometries for progressively larger n. Searching was continued until each of the eight independent basin-hopping simulations had found the same common lowest energy structure. If after ≈1.5 million basin-hopping steps, any of the independent

searches had failed to locate the common lowest energy minimum, the searching with that method was stopped. The method which exhibited full concurrency between all eight independent searches for the largest number of water molecules was considered to be the most efficient global optimisation strategy.

To ensure that the two reference methods had reasonable search parameters across the full range of n considered, the size of the translational step and the temperature were tuned. This tuning was achieved by running 160 000 basin-hopping steps for each parameter combination for every tenth cluster size starting from $n = 5$ (i.e. $n = 5, 15, 25, \ldots$). The combination of temperature and translational step size that consistently found the lowest-energy minima was considered to be optimal, and was assumed to be transferable to neighbouring sites.

For the first reference method (simultaneous rotations and translations, no cycle inversions, constant step size and temperature), concurrency between independent runs was maintained up to $n = 13$ only. In the second method, which introduced cycle inversion moves and organised the steps into blocks, concurrency was maintained up to 18 water molecules. For the adaptive stepsize-and-temperature algorithm, concurrency of searches up to $n = 22$ was observed, indicating that it is the most capable of the three methods investigated for searching hydrated ion landscapes. Although this analysis is far from complete (i.e. it does not investigate the effect of changing search parameters such as the target values or total acceptance ratio) it strongly suggests that the adaptive stepsize-and-temperature algorithm is a sensible and efficient search strategy for such systems.

Chapter 5

Global Optimisation and Energy Landscapes of Hydrated Sulfate Clusters

The sulfate dianion sits at the far kosmotropic end of the Hofmeister series, and its solvation has been studied extensively by experiment [141, 136, 137] and theory [177, 140, 192, 178, 139], due to its relevance in understanding the ordering behaviour, as well as atmospheric chemistry [135], and a number of industrial processes.

Infra-red photodissociation (IRPD) spectroscopy of gas-phase hydrated sulfate clusters, $SO_4^{2-}(H_2O)_n$, in the size range $37 \leq n \leq 80$ has revealed a size dependence of the appearance of OH bonds protruding from the surface of the clusters. The experimental signature, a small peak in the IRPD spectra around 3700 cm^{-1}, is only observed for clusters with $n > 43$, indicating that for smaller clusters, all OH bonds participate in a OH···O hydrogen bond [141]. This result contrasts with the surface of bulk water [193] and small water clusters [173], where many water molecules are able to orient themselves to expose a dangling OH bond, and suggests that the sulfate ion has a significant effect on the surrounding water structure, even beyond the second solvation shell.

Here we present a systematic computational study of $SO_4^{2-}(H_2O)_n$, in the range $3 \leq n \leq 50$, providing microscopic insight into the hydrogen bonding network of water molecules around

the sulfate dianion and into the size-dependent emergence of dangling OH bonds. For clusters with $n \leq 22$, all eight basin-hopping runs located the same lowest energy minimum. For $n = 23$ and 24, five and two of the eight basin-hopping runs located the same lowest energy minimum. These structures are probably good candidate global minima (Figure 5.1). For $n \geq 25$, no two of the eight basin-hopping runs locate a common lowest-energy structure. Therefore, the putative 'global minima' presented for $n \geq 25$ have been observed by only one of the independent basin-hopping searches, and thus they may be representative of low-lying minima, rather than the true global minimum. Complete search statistics and a discussion thereof is presented in Chapter 6. It should be noted that the calculations reported here do no include zero-point vibrational energies, and it has been shown that some reordering of the isomer energies can occur when the are considered [194, 140].

Global minima of $SO_4^{2-}(H_2O)_n$ for $n \leq 7$ have been presented in a previous study by Head-Gordon *et al.* [140] using a combined empirical potential/DFT search method. Our global minima for $n = 4$ and 5 agree with the DFT results (minima 4.5.3 and 5.7.3 in reference 15 respectively), and the DFT global minima for $n = 3$ and 6 (minima 3.6.0 and 6.7.5 in reference [140] respectively) are low-energy isomers (but not global minima) for the present potential. We were unable to reproduce the structure previously proposed as the global minimum for $n = 7$ (minimum 7.9.5 in reference [140]). However, when treated at the DFT level using the B3LYP exchange-correlation functional and 6-311++G** basis set, this structure rearranged to a different geometry, previously reported as the third minimum, which we did find in our global optimisation study. Our empirical global minimum for $n = 7$ is also the lowest energy structure when relaxed and evaluated at the DFT level, and corresponds to the second lowest DFT isomer previously reported (minimum 7.8.6 in reference [140]). To the best of our knowledge, global minima for $n \geq 8$ have not been reported in the literature, so for clusters above this size we have no previous results to compare with. We report the results of our reoptimised DFT calculations for $3 \leq n \leq 18$ at the end of this chapter (Section 5.5).

The evolution of the energy per water molecule, $U(n)/n$ (Figure 5.2) suggests that $n = 50$ is not sufficient to have reached the asymptotic limit of the bulk, infinitely dilute, solution.

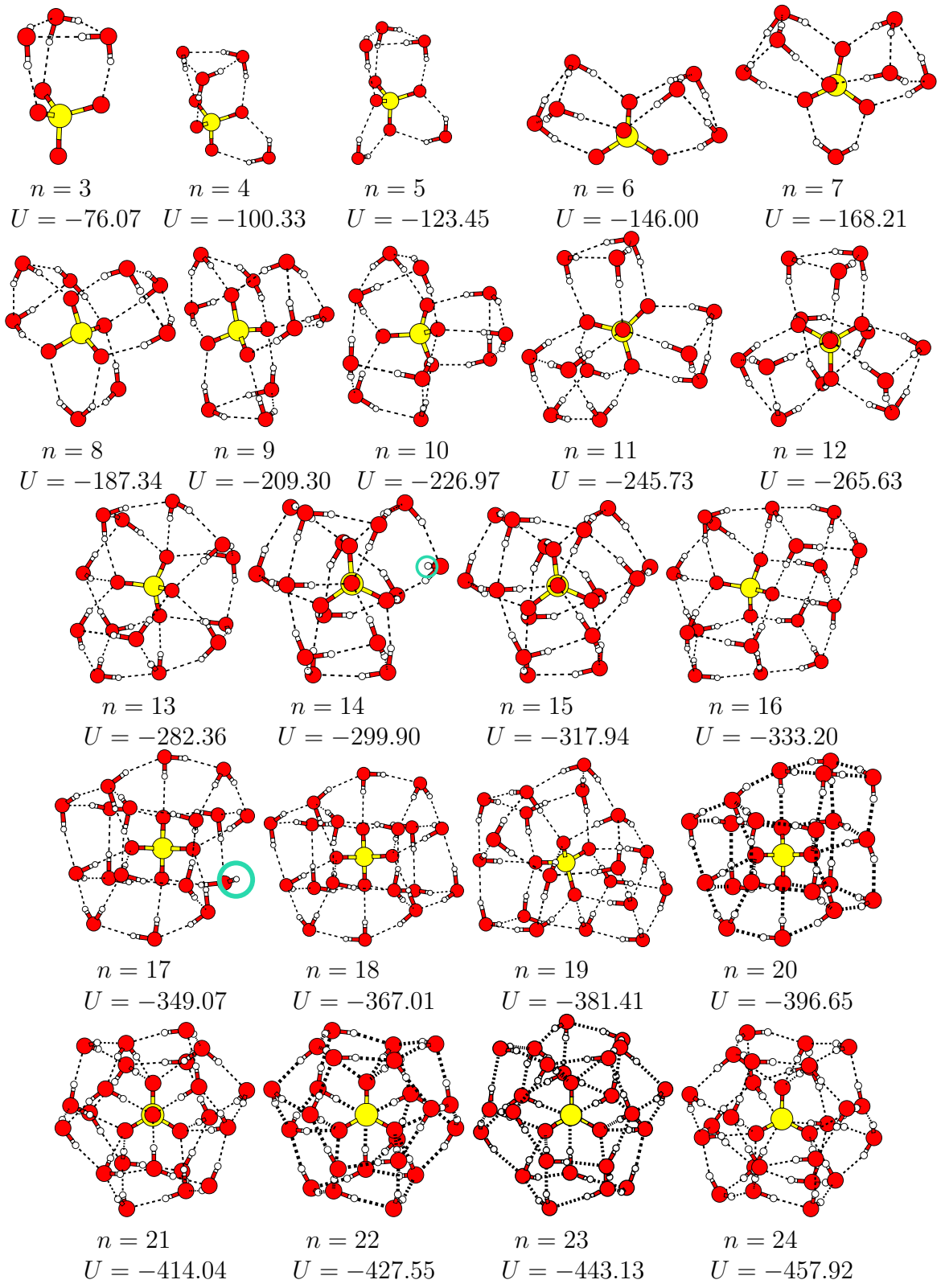


Figure 5.1: Structures and binding energies, U, of the putative global minima for $SO_4^{2-}(H_2O)_n$ clusters where $3 \leq n \leq 24$. The semi-dangling OH bonds in $n = 14$ and 17 are highlighted. Structures of the putative global minima for $25 \leq n \leq 50$ are presented at the end of the Chapter (Figure 5.22). Energies are in kcal mol^{-1}.

We can identify particularly stable cluster sizes using the central difference approximation

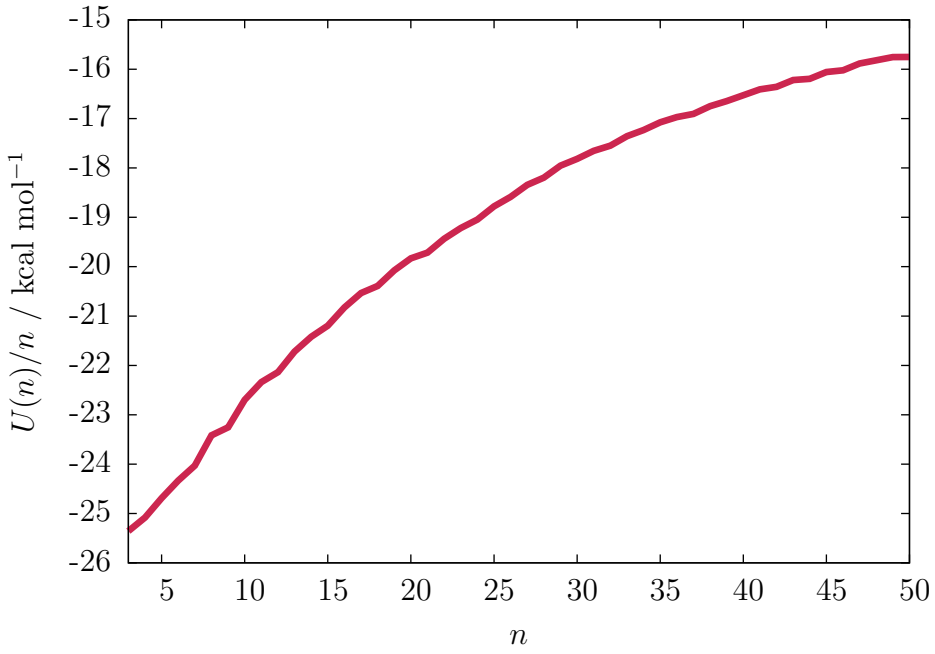


Figure 5.2: Energy per water of the putative global minima of $SO_4^{2-}(H_2O)_n$ clusters, $U(n)/n$, as a function of the number of water molecules, n, for $3 \leq n \leq 50$.

to the second derivative of $U(n)$, (Figure 5.3):

$$\Delta_2 U(n) = \frac{1}{2}\left(U(n+1) + U(n-1)\right) - U(n) \tag{5.1}$$

The more positive $\Delta_2 U(n)$, the more stable that structure is relative to its next nearest neighbouring sizes. In the size range $3 \leq n \leq 14$, the more stable structures are those that contain a high proportion of trimeric water rings ($n = 7, 9, 12, 15$). For $24 \leq n \leq 50$ we observe alternating even-odd behaviour, with structures containing even numbers of water molecules being more stable in most cases. This result contrasts with the behaviour of homogeneous TIP4P water [173], where even-odd behaviour is observed for smaller clusters, before disappearing at larger sizes.

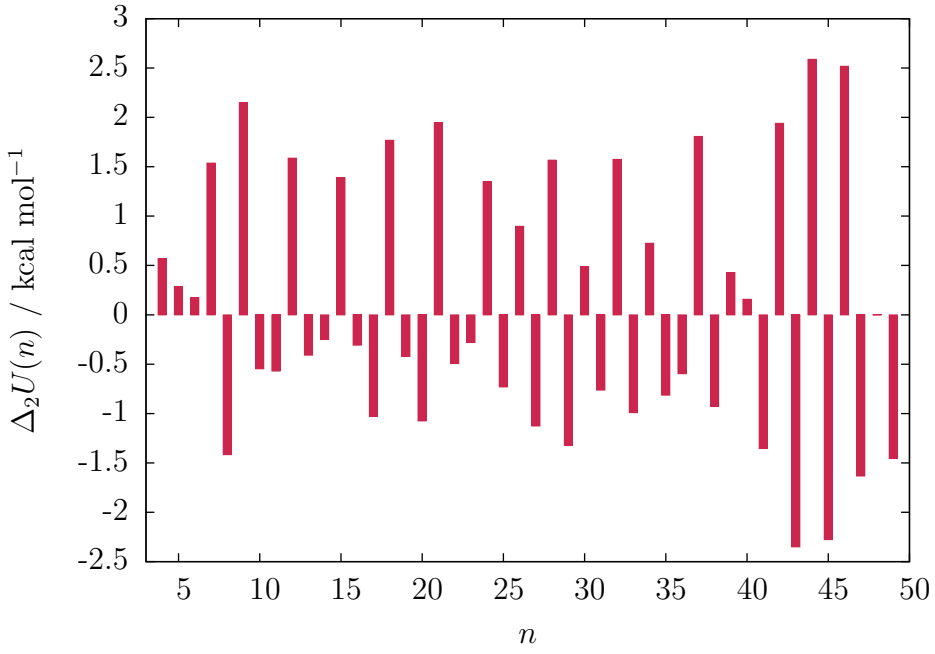


Figure 5.3: Central difference approximation to the second derivative of the energy of the putative global minima, $\Delta_2 U(n)$, as a function of n.

5.1 Analysis of Hydrogen Bonding Behaviour in Low-Lying Minima for $3 \leq n \leq 50$

In Figure 5.4 we plot the mean hydrogen bond length between a pair of water molecules, r_{OH}^{ww}, and between a water molecule and one oxygen of the sulfate ion, r_{OH}^{ws}, of the global minima as a function of n. r_{OH}^{ww} is consistent with the mean hydrogen bond length in pure TIP4P water clusters, and r_{OH}^{ws} is $\approx 4\%$ shorter than r_{OH}^{ww} for all n. An exception to this trend occurs at $n = 5$, where the mean is skewed by the long hydrogen bonds between the two water molecules bonded along the edge of the sulfate. The comparative shortening of hydrogen bonds between water molecules and the sulfate is presumably due to the larger partial charge on the oxygen atom in the sulfate ion than on the lone pair in the TIP4P water. The shortening of r_{OH}^{ws} from $n = 13$–25 is probably due to the compression of the inner water shell in order to allow for tangential hydrogen bonding between water molecules that might otherwise be too far apart. Fluctuations between 12–14 coordination of the sulfate ion is responsible for the oscillatory behaviour in r_{OH}^{ws} above $n = 25$. Here, we define the length of a hydrogen bond to be the OH···O distance. In order to study the size-dependent

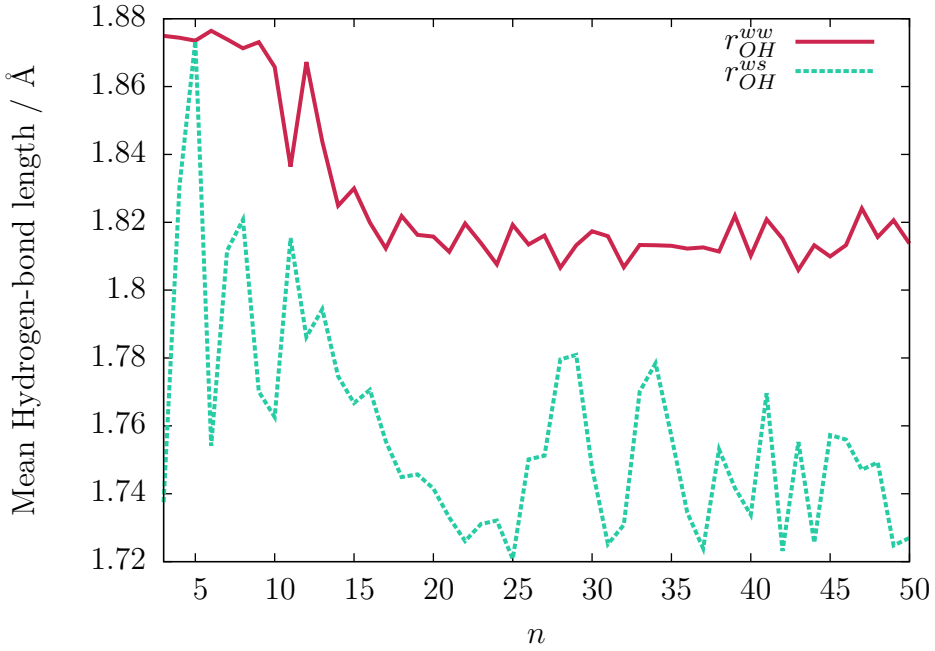


Figure 5.4: Mean hydrogen bond length between water molecules, r_{OH}^{ww}, and between water and the sulfate ion, r_{OH}^{ws}, for the putative global minimum structures of $SO_4^{2-}(H_2O)_n$ clusters as a function of the number of water molecules, n, shown in red and green, respectively.

appearance of dangling OH bonds we examined the hydrogen bonding network of water molecules and the sulfate ion. We deemed structures with a water molecule donating only a single hydrogen bond, either to another water molecule or to the sulfate ion, to contain a dangling OH bond. Dangling OH bonds protruding radially from the cluster were observed in the global minimum structures for the larger hydrated sulfate clusters at $n = 43$, 45 and 47. In the smaller clusters $n = 14$ and 17 we observe semi-dangling OH bonds, which do not engage in a hydrogen bond, but do not protrude radially from the cluster surface. Subsequent local minimisations of these structures at the DFT level (B3LYP, 6-311++G**) shows that these semi-dangling OH groups remain and are not an artefact of the potential. We define $\bar{f}$ as the Boltzmann-weighted mean number of dangling OH bonds,

$$\bar{f} = \sum_i \frac{e^{-\Delta U_i \beta} f_i}{e^{-\Delta U_i \beta}} \tag{5.2}$$

where f_i and ΔU_i are the number of dangling OH bonds and energy above the global minimum for minimum i, respectively, and $\beta \equiv 1/k_B T$. The sum is taken over all unique structures found during any one of the independent basin-hopping global optimisation searches

or (for $3 \leq n \leq 12$) in transition state searches. In Figure 5.5 we plot $\bar{f}$ as a function of n, weighted by temperature, $T = 130$ K to be consistent with experiment [141], and assume that at this temperature the free energy can be estimated from the potential energy alone (i.e. that the entropic contribution to the free energy from each isomer is approximately the same). For the size range $3 \leq n \leq 24$ there are three peaks in $\bar{f}$ at $n = 8$, 14 and 17 water

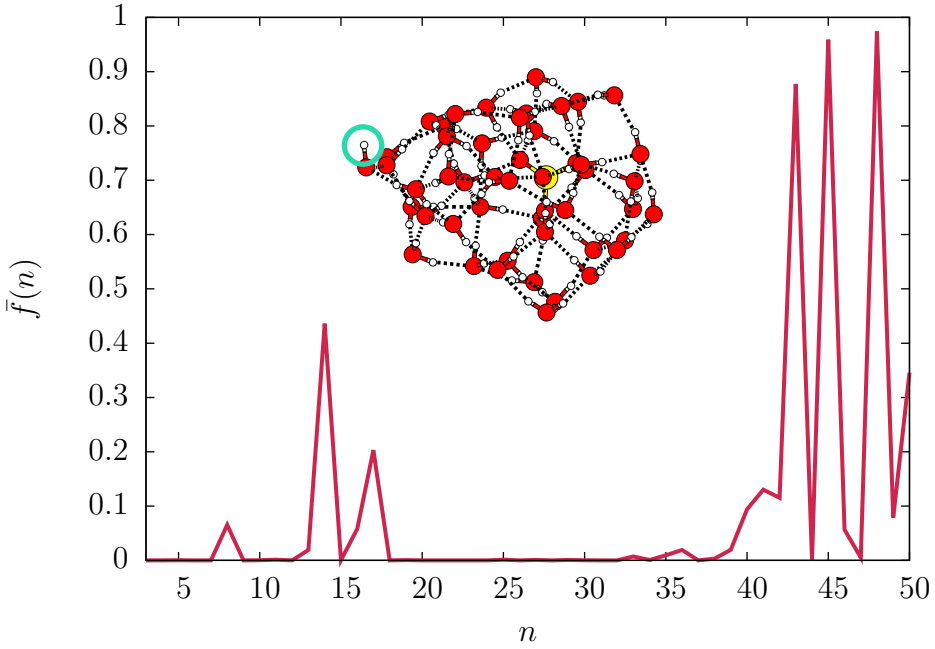


Figure 5.5: Boltzmann-weighted mean number of dangling OH bonds, $\bar{f}$, as a function of the number of water molecules, n, for $3 \leq n \leq 50$. Inset is the global minimum structure for $n = 45$, with a dangling OH bond highlighted.

molecules, due to the global minima of $n = 14$ and 17 exhibiting a dangling OH bond each. For $SO_4^{2-}(H_2O)_8$, there exist a number of low energy isomers with one or two dangling OH bonds, which are stabilised by the other water molecules participating in two trimeric water rings. For $25 \leq n \leq 39$ we observe no energetically relevant structures exhibiting a dangling OH bond. For $40 \leq n \leq 50$ we begin to observe system sizes with prevalent numbers of dangling OH bonds more frequently, with $\bar{f} \approx 0.9$ at $n = 43$, 45 and 48. At this stage, it is not obvious whether the pseudo odd-even behaviour of $\bar{f}$ in this size range is physical, or an artefact due to the difficulty of sampling structures at this size. Provisionally, the data is consistent with the IRPD spectra of size-selected hydrated sulfate clusters [141], which indicate that the sulfate ion suppresses the appearance of dangling OH bonds until $n \approx 43$.

This suppression is not currently well understood, but we can begin to rationalise it as fol-

lows. A water molecule is capable of donating and accepting up to two hydrogen bonds (i.e. four hydrogen bonds in total). In order for a water-containing system to totally inhibit the appearance of dangling OH bonds, each water *must* donate both OH bonds into a hydrogen bond. For systems solely composed of water molecules, this can only be achieved in highly coordinated environments such as bulk ice. At surfaces or in clusters, such coordination numbers are unachievable, and thus dangling OH bonds appear. In hydrated sulfate clusters, the sulfate anion occupies the centre of the cluster, favouring highly coordinated sites. When the first solvation shell is filled, it is able to accept 12–14 hydrogen bonds without donating any back into the system. In this manner, the sulfate anion acts as a net sink for hydrogen bonds, relaxing the requirement that the mean number of donated and accepted hydrogen bonds must be the same. We define $\bar{r}_{ad}$ to be the ratio of the Boltzmann-weighted mean number of hydrogen bonds accepted by and donated to a water molecule,

$$\bar{r}_{ad} = \sum_i \frac{e^{-\Delta U_i \beta} a_i / d_i}{e^{-\Delta U_i \beta}} \tag{5.3}$$

where a_i and d_i are the number of hydrogen bonds accepted by and donated to a water molecule in minima i respectively, and ΔU_i is the energy of minimum i above the global minimum. In Figure 5.6 we plot $\bar{r}_{ad}$ as a function of the number of water molecules n, and observe that for small n, $\bar{r}_{ad}$ is approximately 0.5 (i.e. a water molecule on average donates twice as many hydrogen bonds as it accepts). As the system size increases, the significance of the sulfate ion as a net hydrogen bond acceptor decreases, with $\bar{r}_{ad}$ rising to approximately 0.9 at $n = 50$, implying that 90% of hydrogen bonds are being accepted by water molecules. Note that for systems composed entirely of water molecules, (or in the limit of an infinitely dilute solution), $\bar{r}_{ad} = 1$. Though this analysis provides some insight to the size-dependent inhibition of dangling OH bonds, it is still not clear why they should begin to appear around 40 water molecules and above, and this will be the focus of future work.

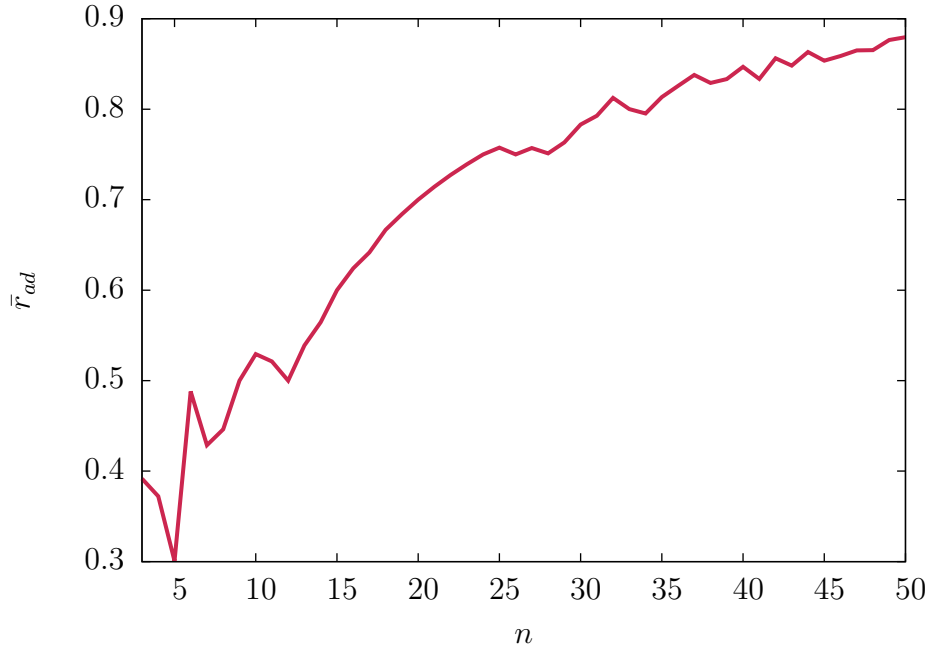


Figure 5.6: Ratio of the Boltzmann-weighted mean number of hydrogen bonds accepted by and donated to a water molecule, $\bar{r}_{ad}$, as a function of the number of water molecules, n, for $3 \leq n \leq 50$.

5.2 Structural Analysis of Global Minima and Energy Landscapes for $3 \leq n \leq 12$

Common structural motifs for the structures of the global minima of $SO_4^{2-}(H_2O)_n$ for $3 \leq n \leq 12$ are trimeric water rings about the open faces of the sulfate ion (see Figure 5.11). In this size range, each putative global minimum structure contains at least one trimer ring, and the structures $n = 3, 6, 9, 12$ have their waters exclusively in such rings. There is some dispute regarding the size-dependent appearance of the second solvation shell, with estimates placed between 8 and 12 water molecules [195, 138]. If we define the second solvation shell as beginning when a water molecule hydrogen bonds only with other water molecules (and not the sulfate ion) we find that the onset is ambiguous. In the lowest energy structures with $n = 10$ and 11, we observe a water molecule which hydrogen bonds only with other water molecules, and participates in a tetrameric ring instead of bonding onto the exposed face of the sulfate ion. At $n = 12$ all water molecules return to sharing a hydrogen bond with the sulfate ion, but for sizes larger than this, at least one water molecule is hydrogen bonded to other waters only.

We built databases of minima and transition states for clusters in the size range $3 \leq n \leq 12$, and present results for $n = 9$ and $n = 12$, which contain 12094 minima and 144677 transition states and 14419 minima and 228415 transition states, respectively, and display interesting kinetic features. Data for other sizes are presented at the end of this Chapter (Figure 5.13). Figures 5.7 and 5.8 show the disconnectivity graph and two lowest-energy isomers of $SO_4^{2-}(H_2O)_9$, respectively. The two isomers both have three trimeric hydrogen bonded water rings (i.e. the same oxygen skeleton), but differ in the relative directionality of the hydrogen bonded rings, resulting in an energy difference of $\Delta U = 0.22$ kcal mol^{-1}. In spite of the energetic and structural similarity, the two isomers are separated by a barrier of ≈ 6 kcal mol^{-1}, leading to a frustrated landscape. The relationship between low-energy isomers that share an oxygen skeleton but differ in hydrogen bond directionality is observed for clusters of other sizes. The disconnectivity graph for $SO_4^{2-}(H_2O)_{12}$ is shown in Figure 5.9, with the five lowest-energy isomers labelled and their structures shown in Figure 5.10. The isomers are organised into two oxygen skeletons, differing in the relative directionality of the hydrogen bonding in the rings alone. Isomers a, d and e have four trimeric hydrogen bonded water rings, and isomers b and c have structures with one trimeric ring and a nine-membered water cycle. Both oxygen skeletons contain 12 water-sulfate and 12 water-water hydrogen bonds. As with the $n = 9$ system, large energetic barriers exist between the isomers, which will lead to frustrated kinetics. In both systems, the energetic gap (≈ 2.5 kcal mol^{-1} and ≈ 1.5 kcal mol^{-1} for $n = 9$ and $n = 12$, respectively) between the labelled minima and the next lowest-energy isomer suggests that the oxygen skeleton geometry is the key factor in determining the energy of hydrated clusters in this size range.

5.3 Structural Analysis of Global Minima for $13 \leq n \leq 24$

Above 13 water molecules, clusters begin to adopt structures containing three-fold symmetric caps of six water molecules about the vertex of the sulfate (three water molecules hydrogen

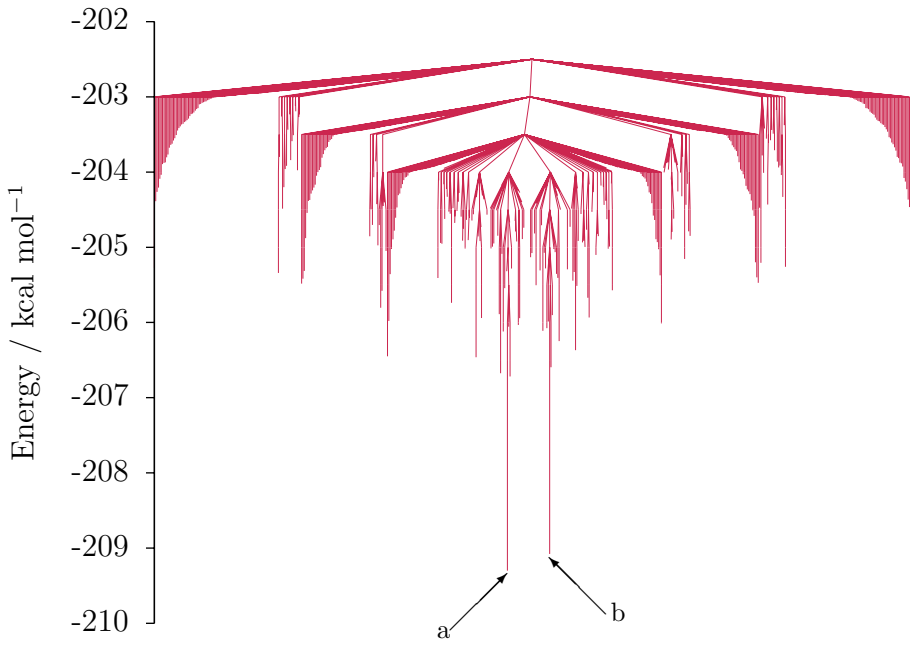


Figure 5.7: Disconnectivity graph for $SO_4^{2-}(H_2O)_9$ containing 555 minima and 1182 transition states connected to the global minimum below -202.5 kcal mol^{-1}.

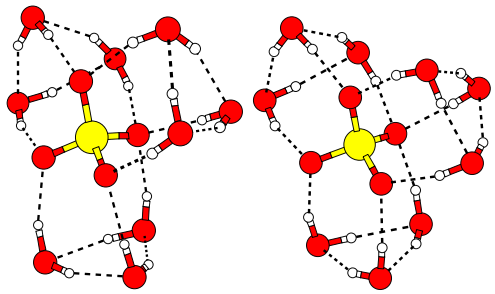


Figure 5.8: The $SO_4^{2-}(H_2O)_9$ global minimum (a) and the next lowest-energy isomer (b), as labelled in the disconnectivity graph in 5.7. ΔU is the energy of a given isomer above the global minimum. Energies are in kcal mol^{-1}.

bond to an oxygen atom on the sulfate and the other three hydrogen bonds between them, see Figure 5.11). These six-membered caps are analogous to trimeric rings for smaller cluster sizes, with each global minimum in this size range containing at least one six-membered cap (with the exception of $n = 13, 16$ and 19). One, two, three and four six-membered caps are observed at every third cluster size from $n = 15$ to 24. When two / three six-membered caps meet, a square / hexagon forms about the intersecting edge / face of the sulfate ion (see Figure 5.11). Clusters containing $n = 14$ and $n = 17$ waters feature a single semi-dangling OH bond, and are the only two global minimum structures found in our study

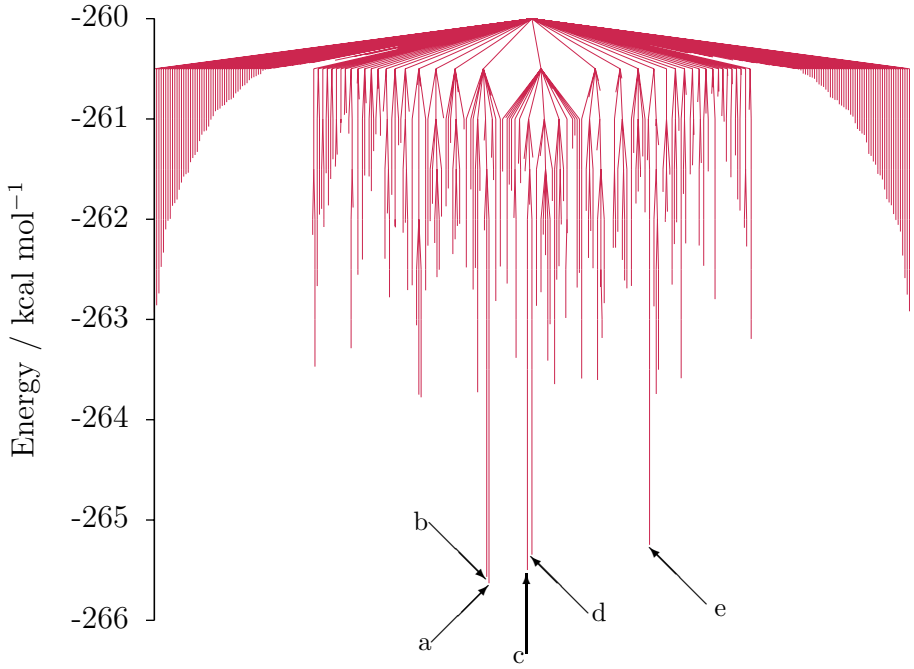


Figure 5.9: Disconnectivity graph for $SO_4^{2-}(H_2O)_{12}$ containing 335 minima and 390 transition states connected to the global minimum below -260.0 kcal mol^{-1}.

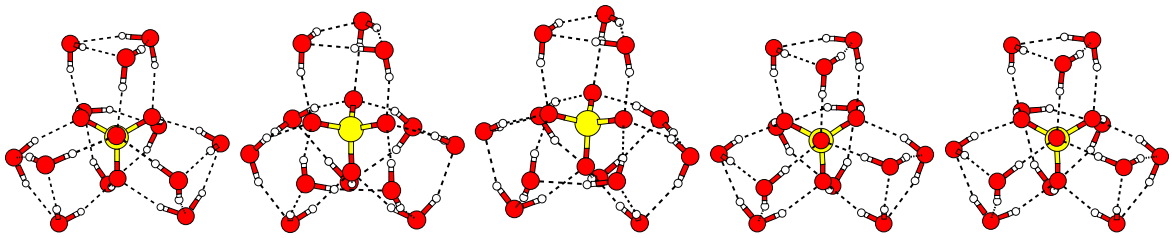


Figure 5.10: The $SO_4^{2-}(H_2O)_{12}$ global minimum (a) and the next four lowest-energy isomers (b-e), as labelled on the disconnectivity graph in Figure 5.9. ΔU is the energy of a given isomer above the global minimum. Energies are in kcal mol^{-1}.

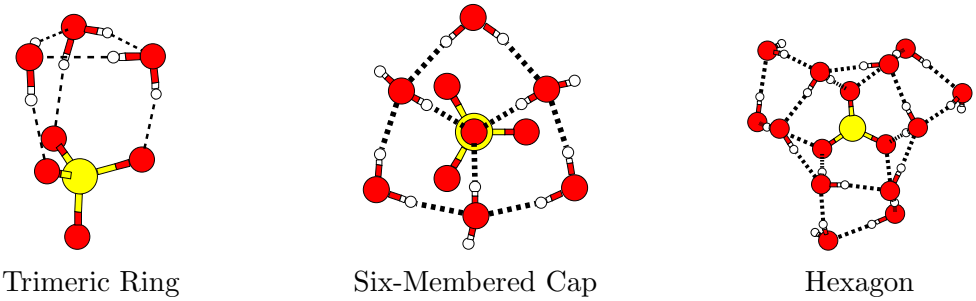


Figure 5.11: Common structural motifs for the $SO_4^{2-}(H_2O)_n$ clusters for $3 \leq n \leq 24$.

below $n = 43$ to do so. In both cases, the structures are similar to their $n + 1$ neighbouring global minimum, but with a water molecule missing from the six-membered cap, causing one of the water molecules to have a dangling OH bond. It should be noted that the dangling OH

bond belongs to a water molecule in the first solvation shell, and which donates a hydrogen bond directly to the sulfate ion. Analysis of the normal mode frequencies of these unusual structures confirm that they are minima on their respective potential energy surfaces.

The largest cluster size at which a trimeric water-ring appears around the open face of the sulfate ion is $n = 15$. For $n \geq 16$, (with the exception of $n = 17$ which has a dangling water), the graph of hydrogen bonded water molecules is connected, there are no self-contained sub-networks of water molecules that hydrogen bond to either themselves or to the sulfate (for example, the four trimeric water rings in $n = 12$ are each disconnected from the other). For clusters containing upto 24 water molecules, broad structural trends in the global minima can be observed (trimeric rings and six-membered caps). The exception to this are the global minimum structures for $n = 13, 16$ and $n = 19$, which exhibit few obvious features or symmetries (though do containing water 'cubes' [172, 196, 197] which share an edge with the sulfate ion). The largest cluster for which the same putative global minimum is found consistently by the majority of independent basin-hopping searches is $n = 23$, with a minority of searches locating a common lowest energy minimum for $n = 24$.

5.4 Structural Analysis of Hydrated Sulfate Clusters for $25 \leq n \leq 50$

For each cluster with $n \geq 25$, the putative global minimum is found in only a single search, so we are not confident that they are the global minima, but we believe that they are structurally representative of low energy isomers. Structures of putative global minima for $25 \leq n \leq 50$ are presented at the end of the Chapter (Figure 5.22). Minima containing one, two or three six-membered caps are a recurring feature, along with structures that appear for the homogeneous TIP4P water clusters, including fused cubes and edge-sharing pentagonal prisms [160]. The sulfate remains fairly central within the cluster throughout. The most important feature is that we do *not* see the emergence of dangling OH bonds in any of the global minima or low-energy isomers until $n \approx 40$, with $\bar{f}$ growing from 0 to ≈ 0.9 (Figure 5.5). The suppression of dangling OH bonds, and the cluster size at which they

begin to appear, is roughly equivalent to what is found in the experimental IRPD spectra.

5.5 Structural Analysis of DFT Global Minima for $3 \leq n \leq 18$

Quantum chemical methods offer a more faithful representation of an atomic or molecular system than an empirical potential, but require significantly more computational resources to compute energies and energy gradients. For this reason, directly searching for low energy minima on a potential energy surface at the quantum chemical level of theory is only possible for small or constrained systems [198, 199, 200]. One can compromise between speed and accuracy by first locating minima on a landscape calculated at a 'lower' level of theory (i.e. using an empirical potential) and then using the lowest energy minima found as a starting point for geometry optimisation at the higher level of theory. Providing there is sufficient agreement between higher and lower levels of theory, such guided searching should return a good sampling of low energy minima at the quantum chemical level, whilst keeping the number of (often very expensive) geometry optimisations to a minimum. In this spirit, the N lowest energy rigid-body minima in the range $3 \leq n \leq 18$ were reoptimised at the DFT level, with $N = 100$ for $3 \leq n \leq 13$ and $N = 40$ for $14 \leq n \leq 18$. The B3LYP exchange-correlation functional and 6-311++G** basis set was used here and in all further DFT calculations presented in this thesis. The rationale for using this particular exchange-correlation functional and basis set is two-fold. Firstly, the reoptimised minima can be compared directly with the results of Head-Gordon *et al.*, who used the same functional and basis set to study hydrated sulfate clusters containing up to seven water molecules (see beginning of chapter for further discussion). Secondly, a comparative study of the treatment of small hydrated sulfate clusters by 24 different density functionals has been reported by Mardirossian *et al.* [192]. Using CCSD(T) calculations as a benchmark, the relative energies and binding energies of a number of $SO_4^{2-}(H_2O)_n$ geometries were calculated using each functional in the size range $3 \leq n \leq 6$. Although B3LYP was comparatively poor at reproducing the binding energies of hydrated sulfate clusters (double hybrid density

functionals were the most successful in this regard), it was capable of calculating the relative energies of different isomers in good agreement of the CCSD(T) results.

The results are shown in Figure 5.12. Each structure is accompanied by three labels; n, the number of water molecules, m, the energetic ordering of the rigid-body minimum from which it was reoptimised, and ΔU, the difference in energy between the minimum in question and the rigid-body global minimum having been reoptimised and evaluated at the DFT level.

There is good agreement between the rigid-body and DFT levels of theory, with nine of the DFT global minima being themselves either empirical global minima or the second or third lowest energy isomer (see correlation plots in Figure 5.13). However, this agreement diverges in the mid-size range ($6 \leq n \leq 13$), where comparatively high-energy rigid-body minima become energetically competitive at the DFT level. At ten water molecules, the 79^{th} lowest rigid-body isomer optimises to the DFT global minimum, though the difference in energy between the empirical and DFT global minima re-optimised at the DFT level of theory, ΔU = 1.54 kcal mol^{-1}, is relatively small. The proliferation of four-membered water rings in global minimum structures for $n \geq 6$ at the expense of trimeric rings suggests that the later is over-stabilised by the empirical potential. For the $n = 12$ structure, the four trimeric rings observed for the empirical global minimum have been replaced by three four-membered rings, and the energy difference between the two DFT-optimised structures is 7.00 kcal mol $^{-1}$. The contrast in ΔU reported for the $n = 10$ and 12 clusters can probably be attributed to how the number of four-membered rings change from the empirical to the DFT global minima: $n = 10$ only gains one such ring, whilst $n = 12$ gains three. The stability conferred at the DFT level by four-membered water rings is such that semi-dangling OH bonds are observed for $n = 8$ and 11, which are stabilised by one and two four-membered rings, respectively. Two semi-dangling OH bonds were observed in the rigid-body global minima for $n = 14$ and 17. These structures still exist at the DFT level, but are no longer global minima, instead being replaced by structures with increased numbers of four-membered rings.

Above $n = 13$ the agreement between empirical and DFT structures improves markedly. It is not clear at this time whether this is a physical observation, or due to the limited number of minimisations performed for clusters containing more than thirteen waters. It

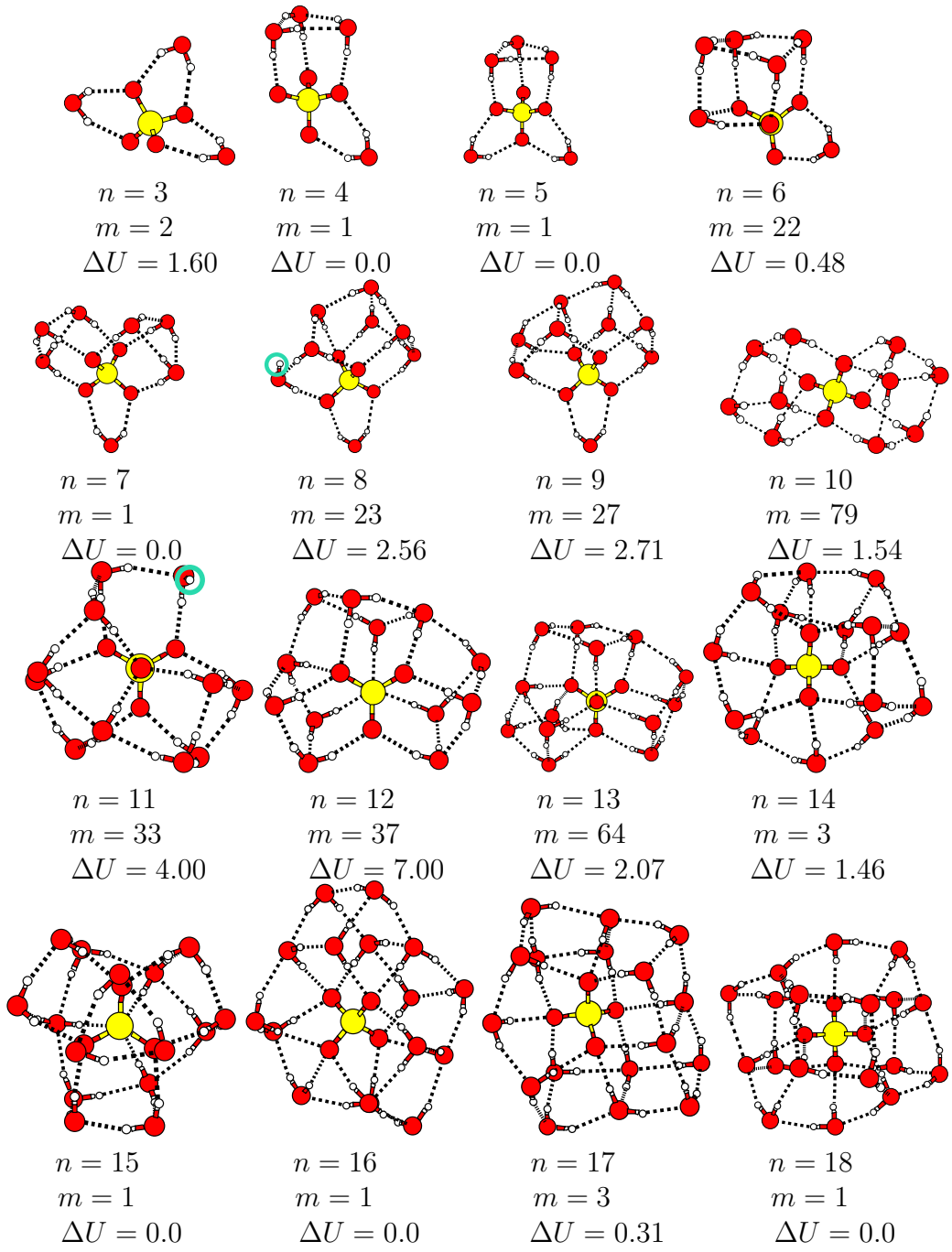


Figure 5.12: Structures of the putative global minima for $SO_4^{2-}(H_2O)_n$ clusters where $3 \leq n \leq 18$, computed a the DFT level using the B3LYP exchange- correlation functional and 6-311++G** basis set. The structures were determined by re-minimising the lowest N structures at each n. For $3 \leq n \leq 13$, $N = 100$ and for $14 \leq n \leq 18$, $N = 40$. The m signifies the energetic ordering of the rigid-body minimum from which the structure was minimised, and ΔU is the difference in energy between the rigid-body and DFT global minimum, when both re-optimised at the DFT level. The semi-dangling OH bonds in $n = 8$ and 11 are highlighted. Energies are in kcal mol^{-1}.

should be noted that the number of desirable structural sub-units such as four-membered rings observed at the empirical level for these sizes could act to stabilise them at the DFT level.

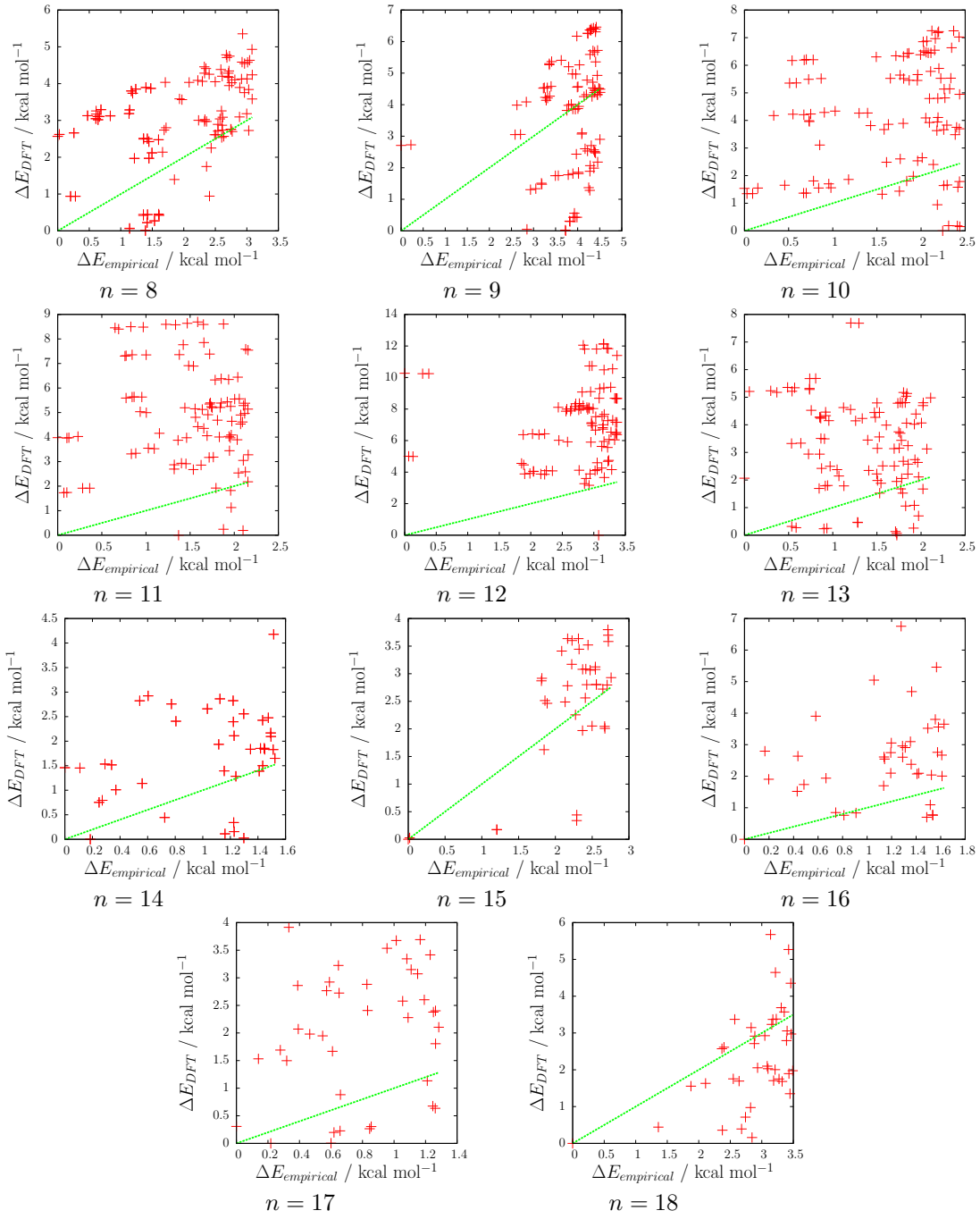


Figure 5.13: Correlation plot of re-optimised DFT minima computed at the B3LYP/6-311++G** level of theory against the original rigid-body minima, for $SO_4^{2-}(H_2O)_n$, with $8 \leq n \leq 18$. Data for $3 \leq n \leq 7$ are presented in Chapter 4.

5.6 Conclusions

The hydrated sulfate ion is an interesting systems both experimentally and computationally for investigating the influence of ion on local water structure. We have shown that rigid-body modelling of hydrated sulfate clusters is capable of replicating the physical chemistry

of these systems. In the size range $3 \leq n \leq 12$, the structures of low-energy minima exhibit water molecules engaging in trimer water rings. The global minima for $n = 3$, 6, 9 and 12 water molecules are particularly stable, with the water molecules participating in such rings exclusively. For these systems, the lowest energy isomers typically have the same oxygen skeleton, but with different relative directionality of the hydrogen bonds, as explored with *cycle inversion* moves. Landscape analysis reveals that although these structures differ only slightly in energy, the barriers to interconversion can be very large, suggesting that such systems will display frustrated kinetics. At larger cluster sizes we note that dangling OH bonds are unfavourable up to $n \approx 43$, consistent with IRPD spectra of size-selected clusters. We suggest that this is due to the sulfate ion acting as a net acceptor of hydrogen bonds, which allows water molecules in the system to accept only a fraction of the hydrogen bonds that they donate. For clusters containing fewer than 19 water molecules rigid-body modelling is in agreement with higher level DFT calculations, with global minima at the DFT level corresponding to either the global minimum or a low-lying isomers of the empirical potential for the majority of cluster sizes considered. However, the rigid-body potential under-estimates the stabilising effect of four-membered rings, causing significant deviation between the empirical and DFT global minima for sizes in the range $6 \leq n \leq 13$. Future work will involve global optimisation of hydrated sulfate clusters directly at the *ab initio* level using the BCGA-DFT code [198, 199] and a combined DFT/basin-hopping [200] approach.

5.7 Global Minimum Structures and Energy Landscapes

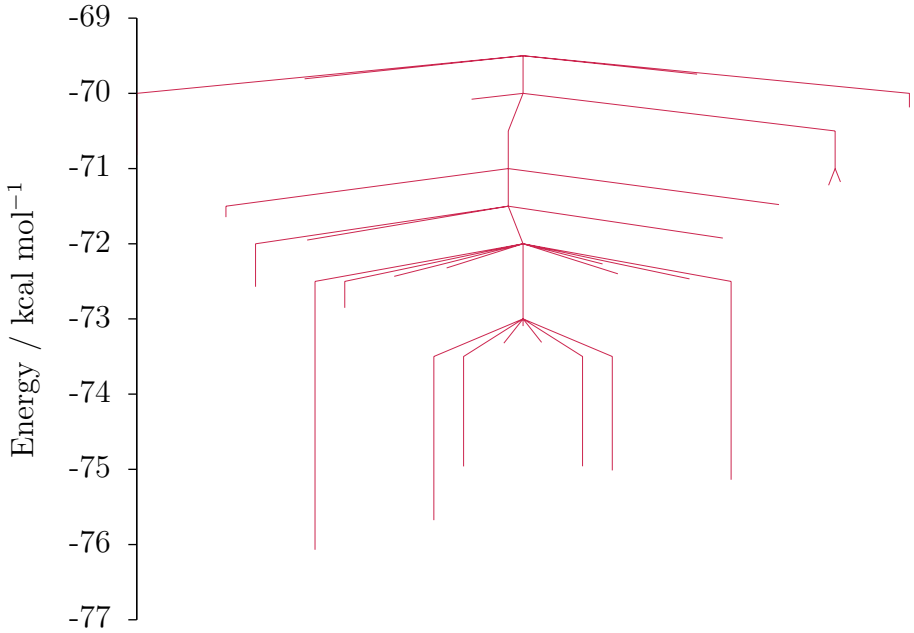


Figure 5.14: Disconnectivity graph for $SO_4^{2-}(H_2O)_3$ containing 27 minima and 98 transition states connected to the global minimum below -69.5 kcal mol^{-1}.

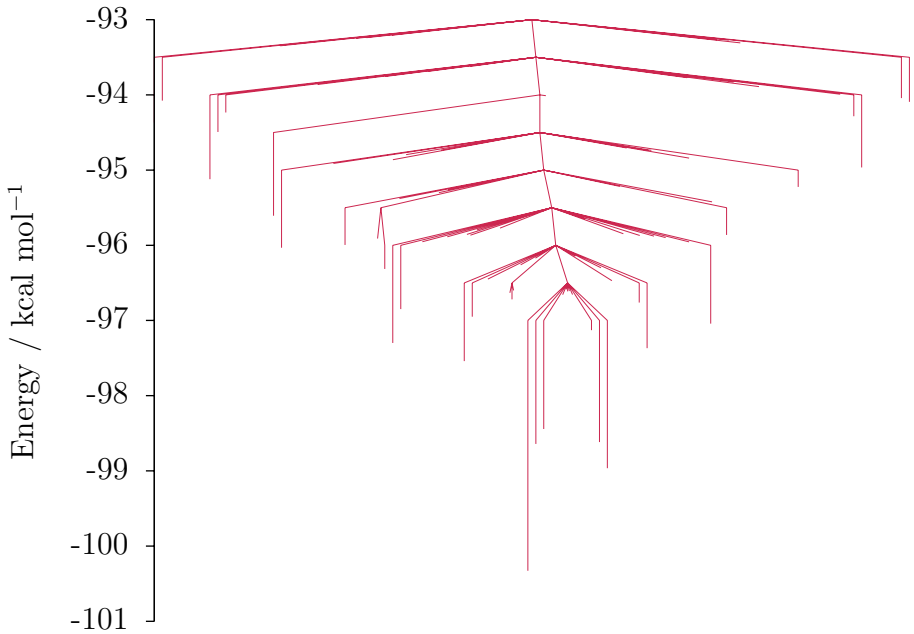


Figure 5.15: Disconnectivity graph for $SO_4^{2-}(H_2O)_4$ containing 96 minima and 427 transition states connected to the global minimum below -93.0 kcal mol^{-1}.

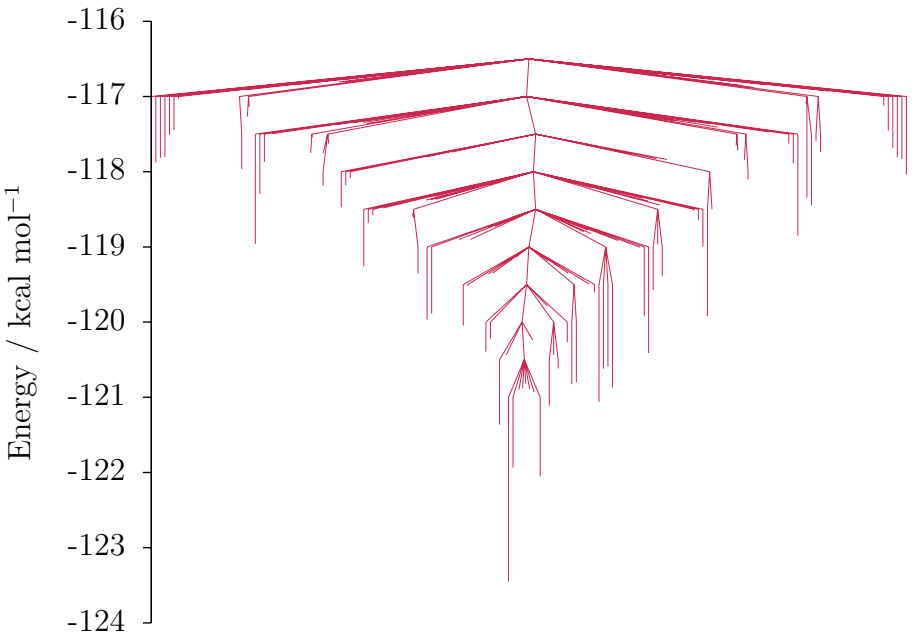


Figure 5.16: Disconnectivity graph for $SO_4^{2-}(H_2O)_5$ containing 168 minima and 581 transition states connected to the global minimum below -116.5 kcal mol^{-1}.

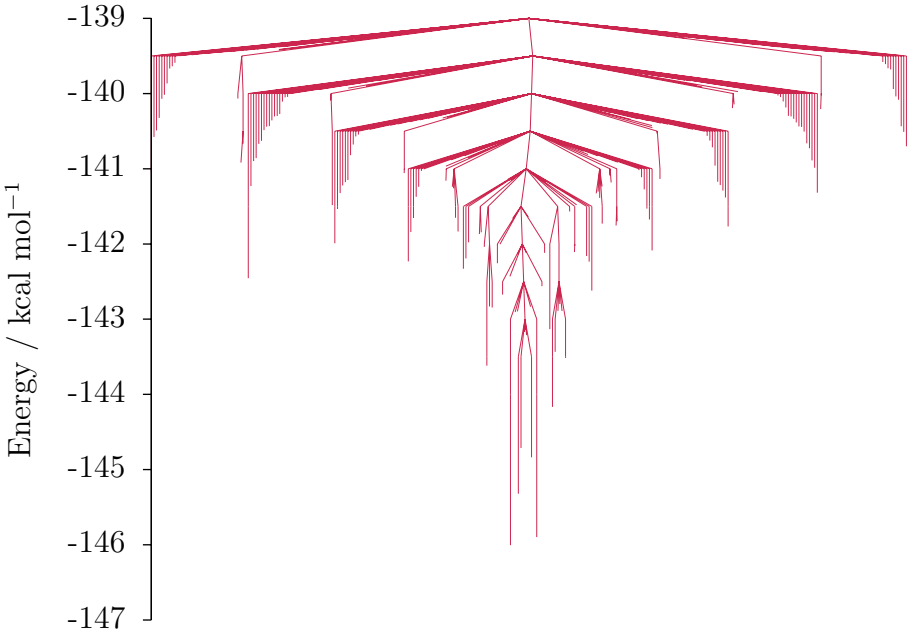


Figure 5.17: Disconnectivity graph for $SO_4^{2-}(H_2O)_6$ containing 295 minima and 1127 transition states connected to the global minimum below -139.0 kcal mol^{-1}.

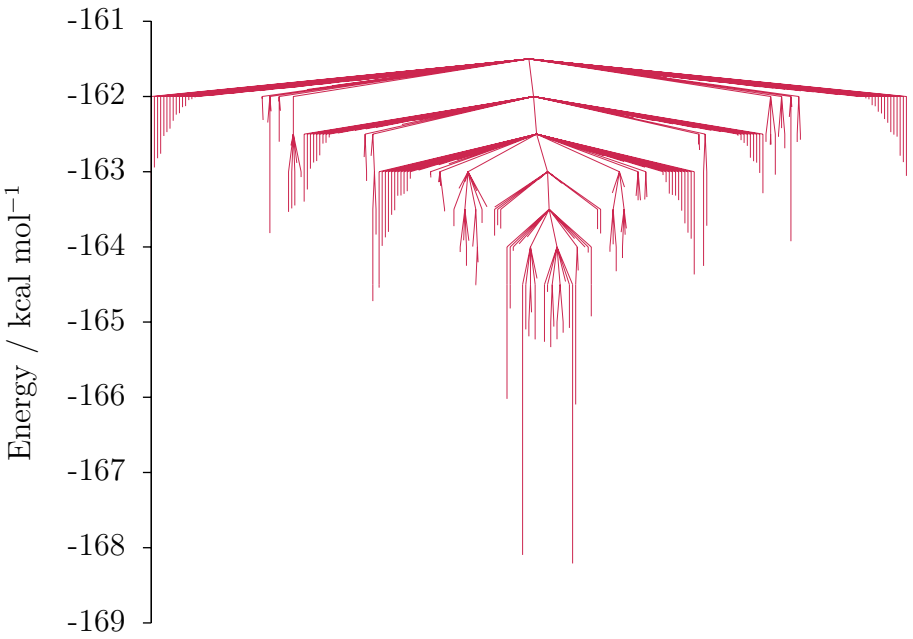


Figure 5.18: Disconnectivity graph for $SO_4^{2-}(H_2O)_7$ containing 243 minima and 602 transition states connected to the global minimum below -161.5 kcal mol^{-1}.

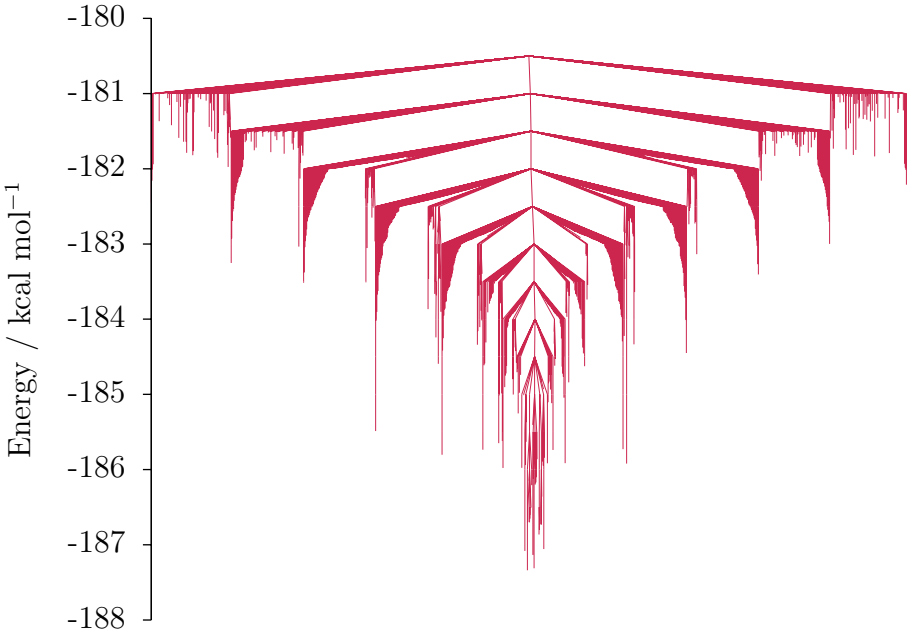


Figure 5.19: Disconnectivity graph for $SO_4^{2-}(H_2O)_8$ containing 1240 minima and 5168 transition states connected to the global minimum below -180.5 kcal mol^{-1}.

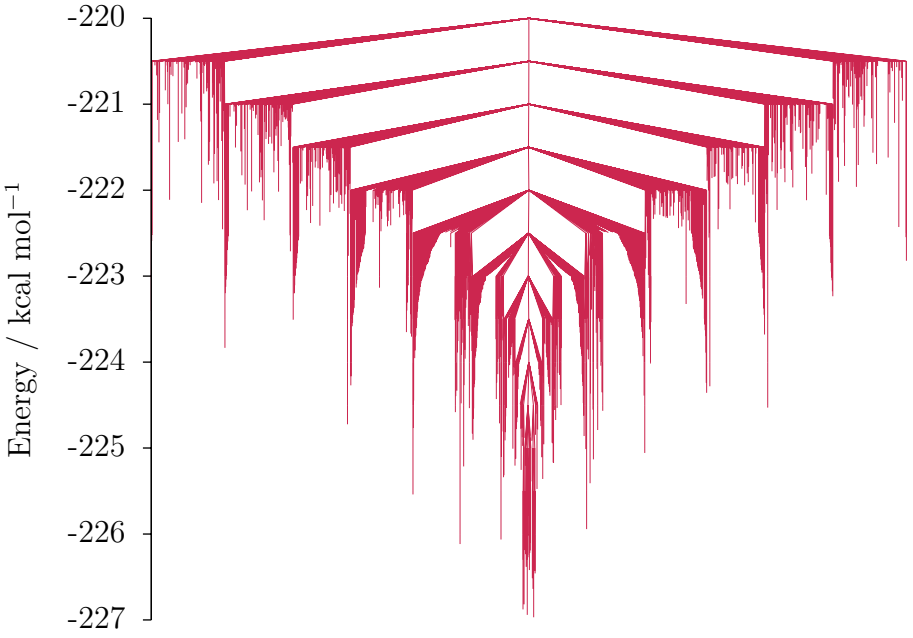


Figure 5.20: Disconnectivity graph for $SO_4^{2-}(H_2O)_{10}$ containing 2102 minima and 9658 transition states connected to the global minimum below -220.0 kcal mol^{-1}.

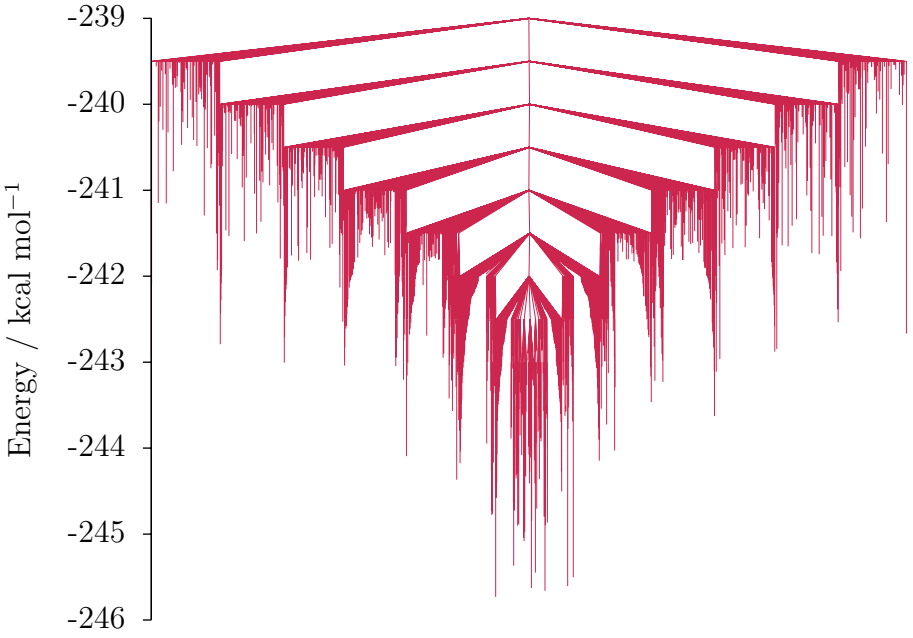


Figure 5.21: Disconnectivity graph for $SO_4^{2-}(H_2O)_{11}$ containing 2268 minima and 9038 transition states connected to the global minimum below -239.0 kcal mol^{-1}.

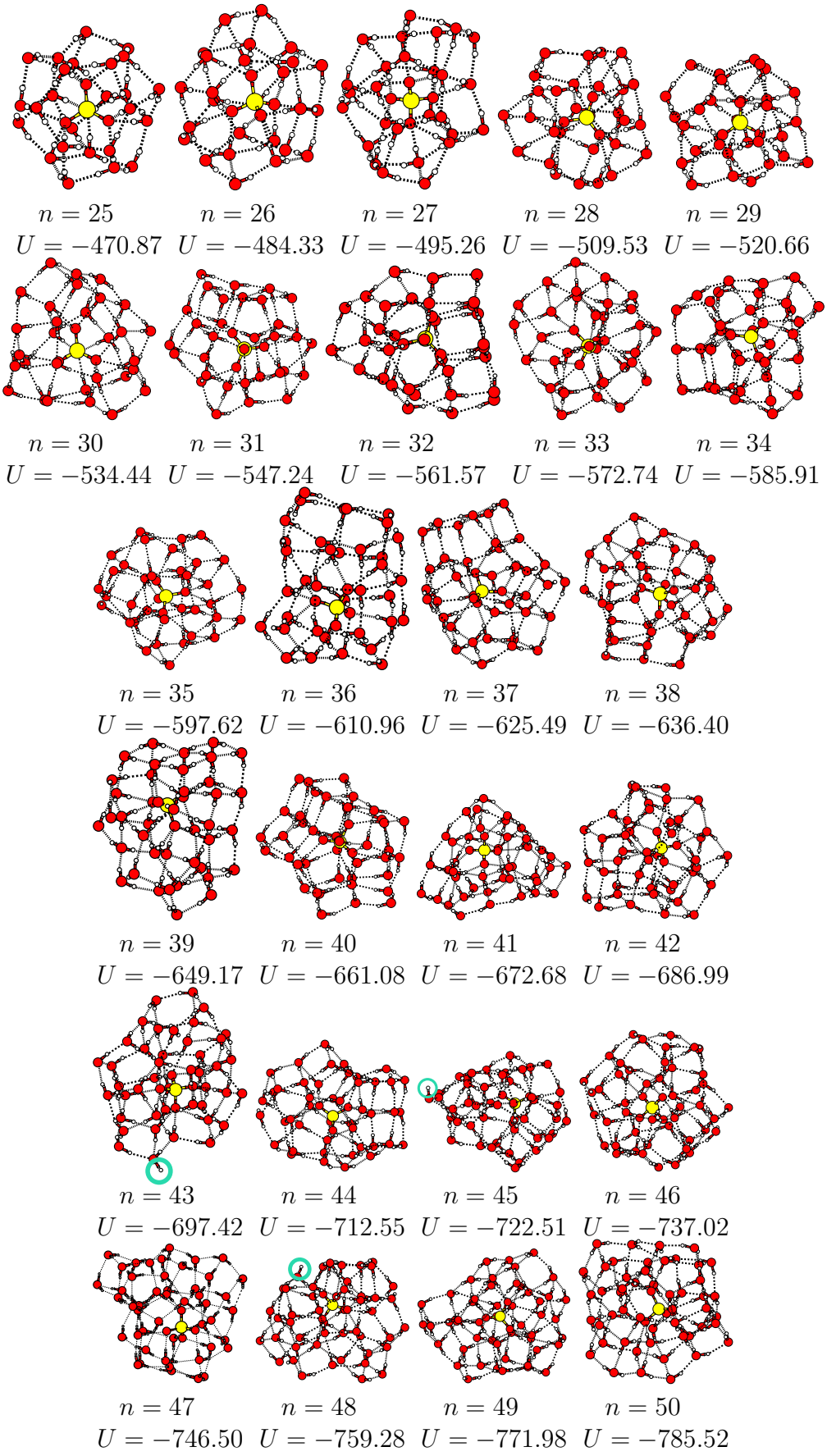


Figure 5.22: Structures and binding energies, U, of the putative global minima for $SO_4^{2-}(H_2O)_n$ clusters where $25 \leq n \leq 50$. The dangling OH bonds in $n = 43$, 45 and 48 are highlighted. Energies are in kcal mol^{-1}.

Chapter 6

Global Optimisation and Energy Landscapes of Hydrated Ion Clusters

In this chapter, the putative global minima and energy landscapes of four hydrated anion systems are presented and analysed. The ions presented are thiocyanate (SCN^-), ferrocyanide ($Fe(CN)_6^{4-}$), ferricyanide ($Fe(CN)_6^{3-}$) and perchlorate (ClO_4^-), which were selected for study either due to their interest to experimentalists or their position in the Hofmeister series [143, 144].

IRPD spectra of gas-phase hydrated ferricyanide ions in the size range $20 \leq n \leq 120$ have been reported in the literature [130] and reveal a size dependence in the appearance of dangling OH bonds around 60-70 water molecules. This extensive suppression of dangling OH bonds indicates that the range and extent to which the ion is capable of influencing its surrounding solvent structure is significant. No experimental data are currently available for size selected hydrated ferrocyanide ions, but we include the results to investigate the effect of charge on the surrounding water molecules.

The thiocyanate and perchlorate ions are both chaotropic anions which sit at the opposite end of the Hofmeister series from the sulfate ion. Their putative global minima and landscapes are reported in this thesis to study the effect on differing ion geometry at constant charge, and as a counterpoint to the chaotropic sulfate ion presented in Chapter 5.

6.1 Structures and Landscapes of Hydrated Thiocyanate Ions

For clusters with $n \leq 19$, all eight basin-hopping runs located the same lowest energy minima. For $n = 20$ and 21, five and two of the eight basin-hopping runs, respectively, located the same lowest common energy minimum. These structures are probably good candidate global minima (Figure 6.3). For $n \geq 23$, no two of the 8 basin-hopping runs locate a common lowest-energy structure, and so the putative 'global minima' presented at this size have been observed by only one of the independent basin-hopping searches. Hence they may be representative low-lying minima, rather than the true global minimum. To the best of our knowledge, putative global minima for hydrated thiocyanate clusters have not been reported in the literature, so we have no previous results to compare with. We report the results of our reoptimised DFT calculations for $3 \leq n \leq 16$ at the end of this section. The central difference approximation shown in Figure 6.2 (and defined by Equation 5.1 in Chapter 5) identifies the global minima for $n = 10$, 16, 22, 24 and 26 as being magic number structures. Each of these structures contains a large number of closed cubic or pentagonal water units, which are common motifs in pure water clusters, and probably confer stability to the structures in question. We observe alternating even-odd behaviour in stability for $n \leq 10$ and $22 \leq n \leq 34$, with structures containing even numbers of water molecules being more stable. For $n \leq 10$, this behaviour correlates with the number of complete four-membered rings contained within the structure. For all sizes studied, the thiocyanate ion prefers to sit on the surface of the cluster, and does not become fully solvated. Common structural motifs for the structures of the global minima of $SCN^-(H_2O)_n$ are four-membered water squares and (above $n = 18$) five-membered water pentagons. These structures are common in the global minima of pure TIP4P water clusters, where four-membered and five-membered water rings are able to stack into cubes and pentagonal prisms of increasing length. We shall use the terminology used by Wales *et al.* [160], and label three or more stacked water rings as fused cubes / fused pentagonal prisms. Structural analogues exist between the global minima of pure TIP4P clusters and hydrated thiocyanate clusters, and comparisons will be drawn

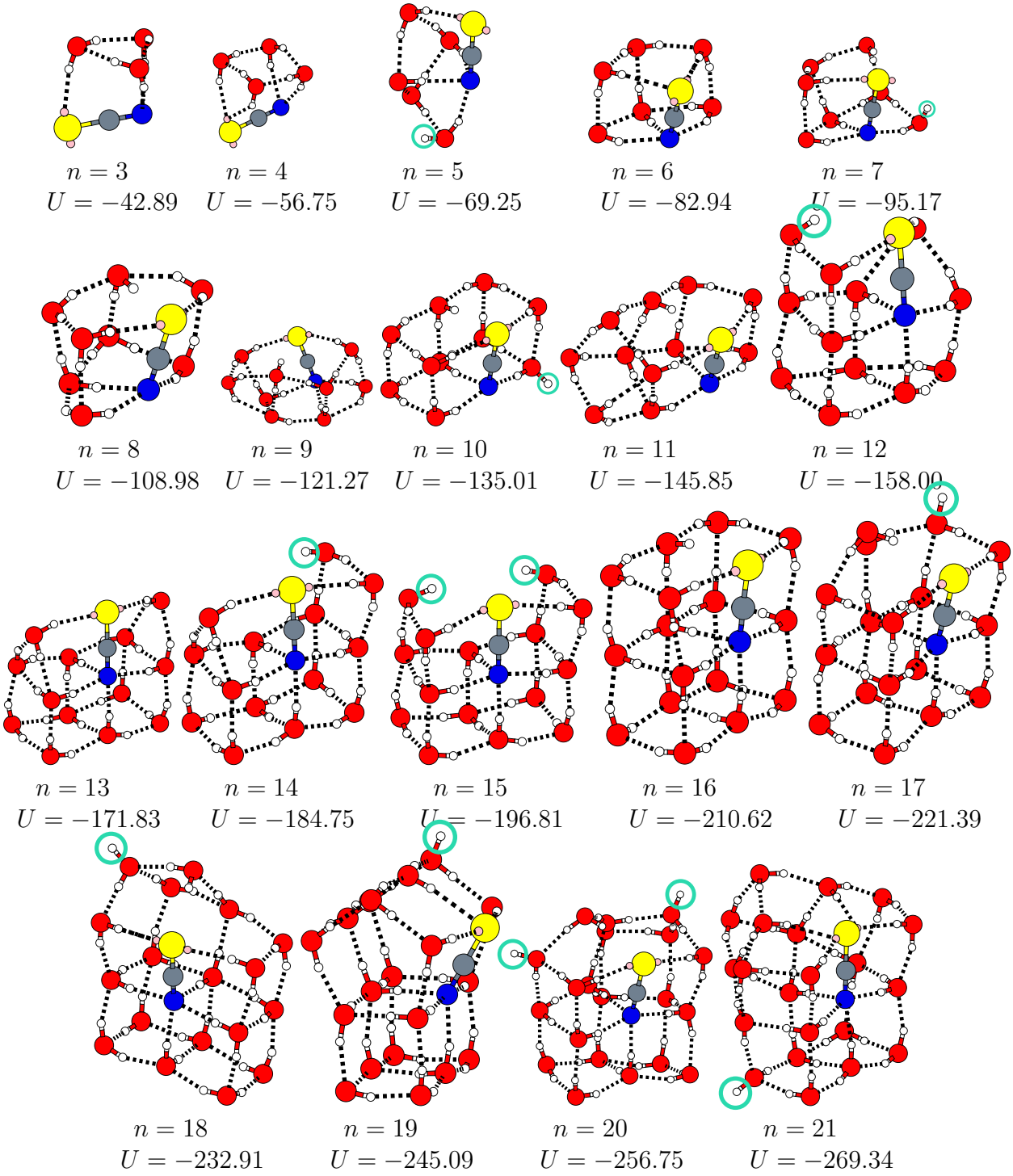


Figure 6.1: Structures and binding energies, U, of the putative global minima for $SCN^-(H_2O)_n$ clusters where $3 \leq n \leq 21$. The dangling OH bonds in $n = 5$, 7, 10, 12, 14, 15, 17, 18, 19, 20 and 21 are highlighted. Energies are in kcal mol^{-1}.

throughout this analysis. The discrepancy in length between the thiocyanate ion (2.84 Å) and the mean OH···O distance (≈ 1.9 Å) forces one end of the ion to protrude from the structure. Invariably, it is the sulfur end of the ion which protrudes, with the nitrogen atom preferring to sit in higher coordination sites than the sulfur (see for example the $n = 13$ struc-

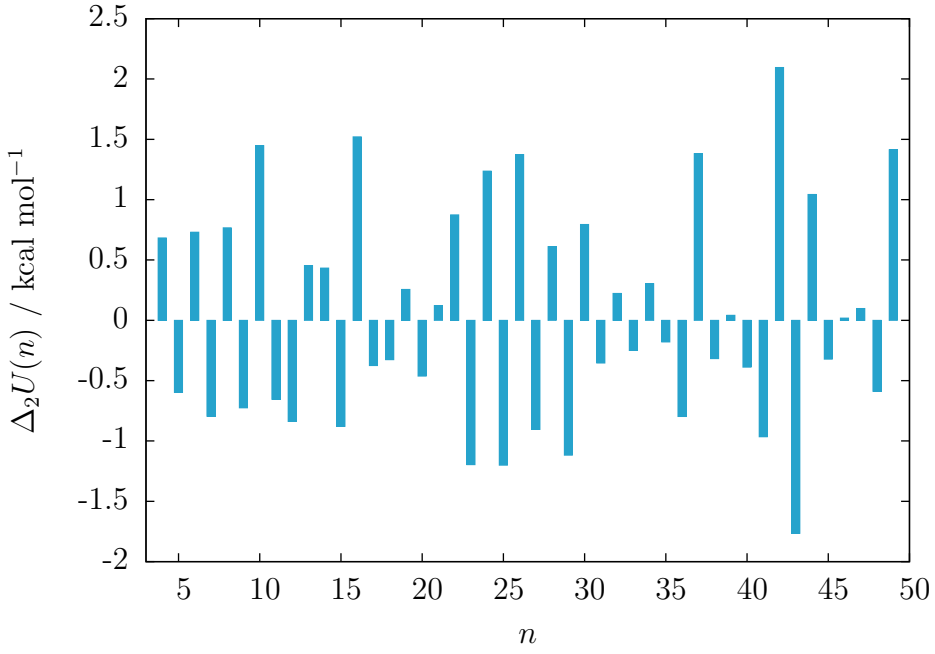


Figure 6.2: Central difference approximation to the second derivative of the energy of the putative hydrated thiocyanate global minima, $\Delta_2 U(n)$, as a function of n.

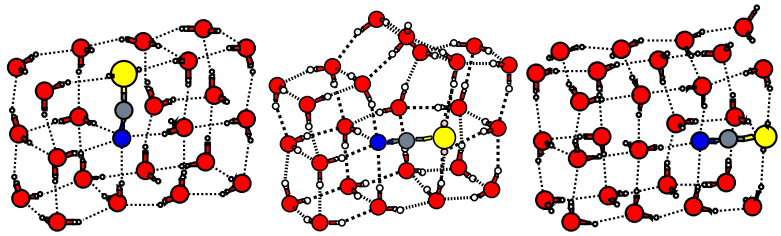


Figure 6.3: Structures and binding energies, U, for $n = 22$, 24 and 26. These structures are of interest due to their relative energetic stability and their geometries, which follow trends observed at smaller n. Energies are in kcal mol^{-1}.

ture, where the nitrogen atom is four- coordinated and the sulfur atom is two-coordinated). In Figure 6.4 the Boltzmann-weighted mean coordination numbers of the nitrogen and sulfur atoms are presented. Of the 47 cluster sizes investigated, only four ($n = 30, 34, 46$ and 50) exhibited a sulfur atom with a higher coordination number than the nitrogen atom. Given the inherent difficulties of sampling large systems, this result is probably more indicative of incomplete searching than any physical phenomenon. The preference for the nitrogen atom to sit at higher coordination sites is consistent with previous results [201], and is likely due to the atom's higher charge density (both the ends of the thiocyanate ion carry a charge of $-0.35e$, but the sulfur atom distributes its charge between a pair of lone pair sites, lowering

the charge density).

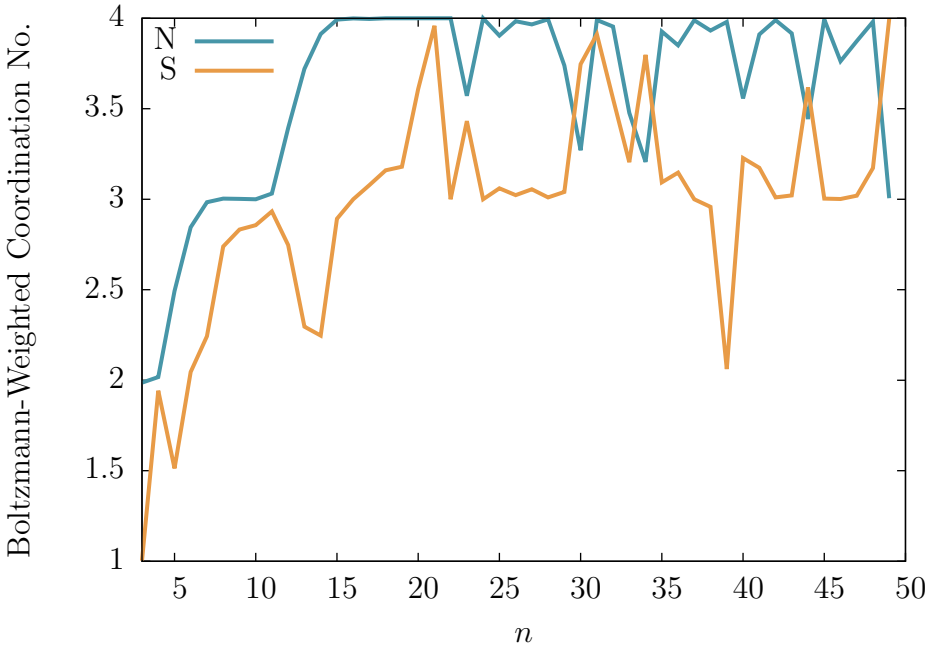


Figure 6.4: Boltzmann-weighted mean coordination number of the nitrogen (blue) and sulfur (yellow) atoms as a function of the number of water molecules, n, for $3 \leq n \leq 50$. Isomers were weighted according to their potential energy, at a temperature of 130 K.

As mentioned in the previous section, water molecules are exclusively engaged in one, two, three and four water squares for $n = 4$, 6, 8 and 10, respectively. The increasing number of water squares culminates in a pair of fused water cubes at $n = 10$, with the SCN^- occupying one of the fused edges, effectively acting as two water molecules. The oxygen-thiocyanate skeleton of $SCN^-(H_2O)_{10}$ is equivalent to the all-oxygen skeleton of the TIP4P $(H_2O)_{12}$ global minimum. For $n \geq 13$, fused water-cubes are observed in all global minimum structures.

Databases of minima and transition states for clusters in the size range $3 \leq n \leq 13$ have been compiled, and results for $n = 10$ and $n = 13$, which contain 10136 minima and 61436 transition states, and 13884 minima and 92880 transition states, respectively, are presented. Figures 6.5 and 6.6 show the disconnectivity graph and four low-energy isomers of $SCN^-(H_2O)_{10}$, respectively. The four isomers include the global minimum (a) and the second (b), sixth (c) and ninth (d) lowest-energy structures, and are representative of the geometric variety observed in the low energy structures of $SCN^-(H_2O)_{10}$. Isomers b and c share an oxygen-thiocyanate skeleton with the global minimum, and just differ by inversions

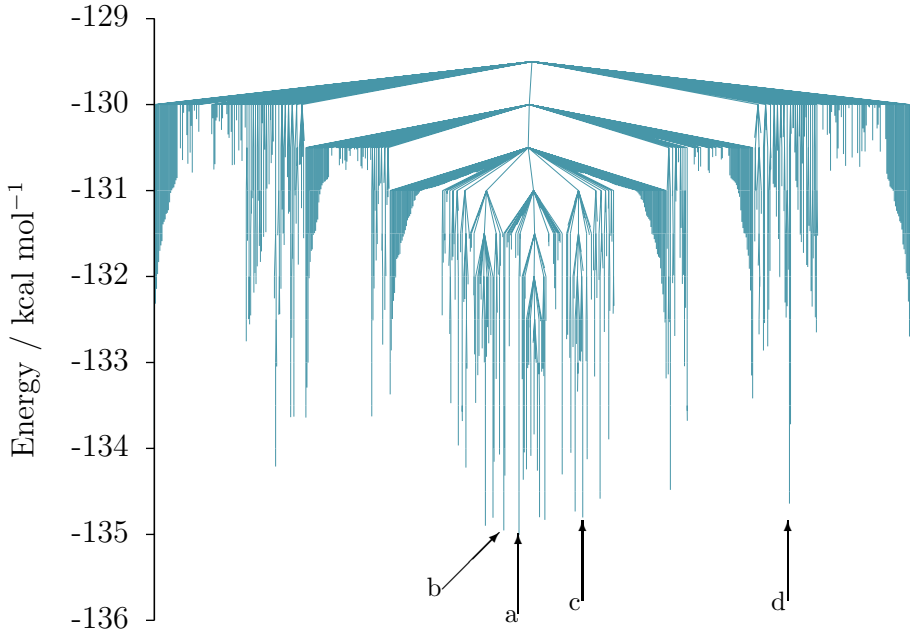


Figure 6.5: Disconnectivity graph for $SCN^-(H_2O)_{10}$ containing 699 minima and 1059 transition states connected to the global minimum below -129.5 kcal mol^{-1}.

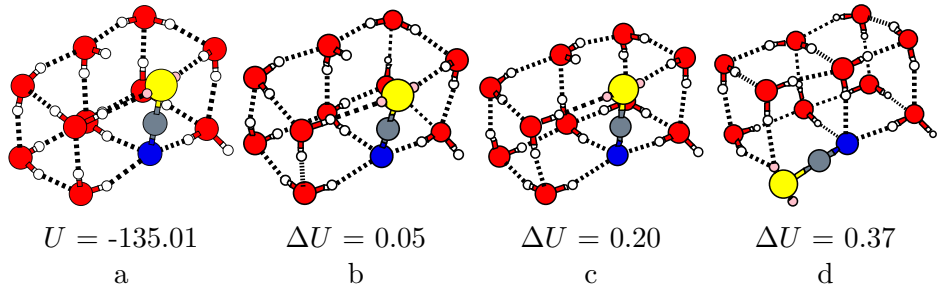


Figure 6.6: The $SCN^-(H_2O)_{10}$ global minimum (a) and the 2^{nd}, 6^{th} and 9^{th} lowest-energy isomers (b-d) respectively, as labelled on the disconnectivity graph in Figure 6.5. ΔU is the energy of a given isomer above the global minimum. Energies are in kcal mol^{-1}.

of the directionality of their hydrogen bond networks (one and two inversions for b and c respectively). Isomer d is still composed of two fused cubes, but with the thiocyanate lying on an external edge leading to a sulfur atom of coordination two. The disconnectivity graph for $SCN^-(H_2O)_{13}$ is shown in Figure 6.7, with four low-energy isomers labelled and their structures shown in Figure 6.8. The isomers are organised into three oxygen skeletons. Isomers a and b (GM and 2^{nd} lowest-energy structures) adopt a 'terraced house' structure, with a fused cube base and triangular prism roof, and differ from each other by a single hydrogen bond cycle inversion. Isomer c is similar to isomers a and b, with the sulfur atom on the thiocyanate ion sitting in a higher coordinate site, but at the expense of disrupting the fused

cube structure. Isomer d has an 'open chest' structure, comprising a fused cube and an open lid containing the thiocyanate ion. This structure demonstrates how an energetically stable sub-structure (in this instance the fused water cube) can stabilise a structure containing otherwise unfavourable components (such as the two under- coordinated water molecules in the lid). Other examples can be observed in the global minimum structures of the $n = 12$, 14 and 15 hydrated thiocyanate clusters.

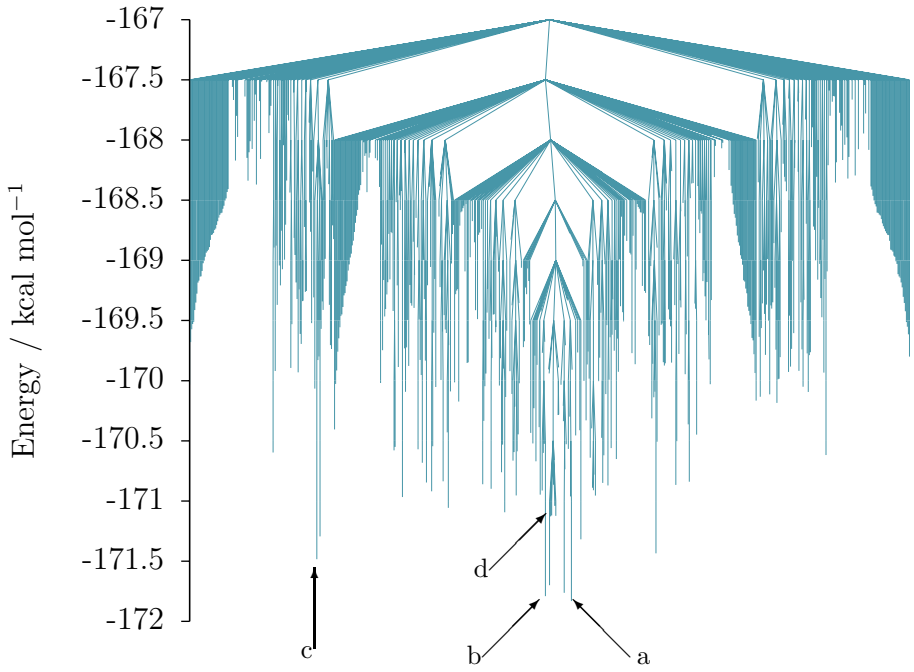


Figure 6.7: Disconnectivity graph for $SCN^-(H_2O)_{13}$ containing 691 minima and 925 transition states connected to the global minimum below -167.0 kcal mol^{-1}.

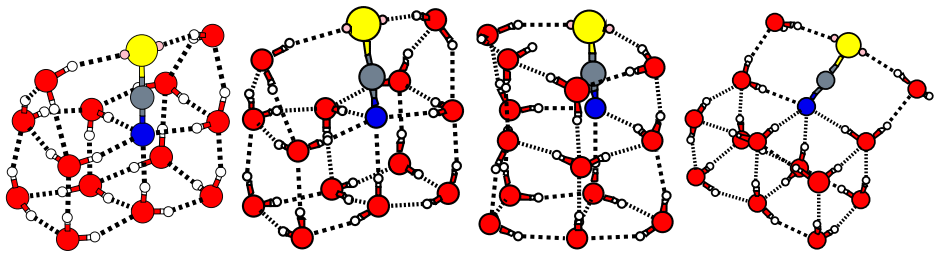


Figure 6.8: The $SCN^-(H_2O)_{13}$ global minimum (a) and the 2^{nd}, 5^{th} and 10^{th} lowest-energy isomers (b-d) respectively, as labelled on the disconnectivity graph in Figure 6.7. ΔU is the energy of a given isomer above the global minimum. Energies are in kcal mol^{-1}.

For both the 10-and 13-water molecule systems, low-energy isomers share a small number of oxygen-thiocyanate skeletons, but differ by the placement of the hydrogen atoms within

the structure (for example, of the lowest 40 minima on the $n = 10$ landscape, all but one structure exhibits a fused cube skeleton). In spite of the structural and energetic similarity of these minima, the barriers between different cycle isomers is large (≈ 4 kcal mol^{-1} for $n = 10$) and larger still when the oxygen-thiocyanate skeleton is altered (≈ 5 kcal mol^{-1} to convert from isomers a and d on the $n = 10$ landscape). Large energetic barriers imply slow relaxation times between minima, and when coupled with energetically competitive isomers suggests that the system will exhibit glassy or frustrated dynamics. As the number of water molecules grows, the number of possible permutations of hydrogen atoms with a given oxygen skeleton increases. Thus, it is reasonable to assume that the landscape will grow *increasingly* frustrated with increasing n.

The $SCN^-(H_2O)_{16}$ global minimum has an oxygen-thiocyanate skeleton organised into a pair of fused cubes (fused along the long face of each), with the nitrogen atom sitting in the central four co-ordinate site, and the sulfur on a three co-ordinate edge site. There is little variation in the oxygen-thiocyanate skeleton amongst the other low-energy isomers (all but one of the lowest 50 energy minimum structures share an oxygen-thiocyanate skeleton). Unlike the $n = 10$ system, isomers in which the thiocyanate ion sits along the edge of a cube (see isomer d in Figure 6.6) are not observed, nor are structures where the sulfur atom sits in the other four co-ordinate site adjacent to the nitrogen atom, solvating the ion. The trend of adding additional fused cubes to a structure continues at $n = 22$: the putative global minimum structure is composed of three fused cubes. Central difference analysis (Figure 6.2) identifies the $n = 10, 16$ and 22 global minimum structures as magic number clusters, due to the stabilising effect of having all water molecules participate in a water-thiocyanate cube.

The recurrent use of stable sub-units to construct larger energetically favourable superstructures is common to both hydrated thiocyanate and pure TIP4P water clusters. Examples composed of the structural motifs observed in both systems are shown and characterised in Table 6.1. As discussed previously, four-membered water squares (which can stack to form cubes, fused cubes etc.) and five-membered water pentagons, which aggregate into pentagonal prisms are the predominant sub-units for both, though the manner in which these units aggregate differs between the two systems. For TIP4P water, the fused-cube motif stacks

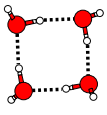
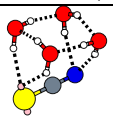
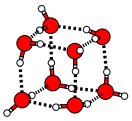
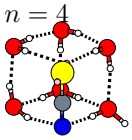
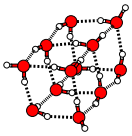
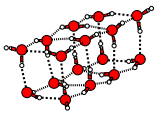
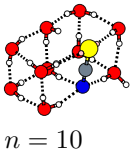
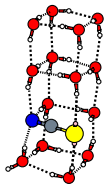
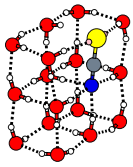
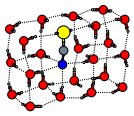
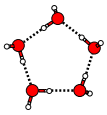
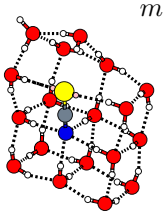
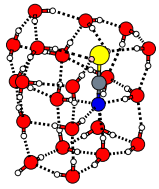
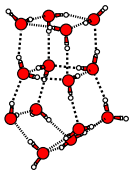
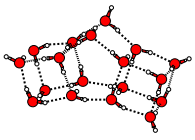
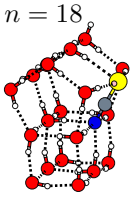
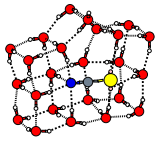
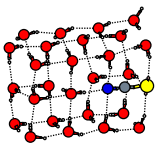
Motif	$(H_2O)_n$	$SCN^-(H_2O)_n$
Square	 $n = 4$	 $n = 4$
Cube	 $n = 8$	 $n = 6$ $m = 5$
Fused-Cube	 $n = 12$  $n = 16$	 $n = 10$  $n = 14$ $m = 10$  $n = 16$  $n = 22$
Pentagon	 $n = 5$	 $n = 18$  $n = 21$
Fused-PC	 $n = 14$  $n = 18$	 $n = 19$  $n = 24$  $n = 26$

Table 6.1: Recurring structural motifs common to both low energy isomers of TIP4P water clusters and hydrated thiocyanate clusters. Structures which are not global minima are labelled with m, which signifies the energetic ordering of the isomer. Fused-PC denotes structures containing fused pentagonal prisms and cubes. TIP4P water structures are taken from the Cambridge Cluster Database [160].

into one-dimensional chains, whilst the hydrated thiocyanate global minima tend to aggregate into two dimensional structures. This difference in dimensionality is probably driven by the energetic benefit of placing the nitrogen atom in a more highly coordinated site within the water structure. Pentagons and pentagonal prisms appear in the global minima of the larger hydrated thiocyanate ions (from $n = 18$ onwards), either as a pentagonal face in an otherwise cubic structure ($n = 18$ and 21) or in a combined fused-pentagonal-prism-and-cube arrangement ($n = 19, 24$ and 26). Lone pentagons or pentagonal prisms (such as those for pure TIP4P water at $n = 5, 10, 15$ etc.) are not observed for $SCN^-(H_2O)_n$ due to the low average coordination number that the nitrogen atom is required to adopt. One can note that the global minimum structures for six and eight waters are related to the single cube and pentagonal prism, respectively, but with the water molecules on adjacent edges to the thiocyanate reorienting themselves to donate a hydrogen bond to the ion.

The twenty lowest energy rigid-body minima for all cluster sizes in the in the range $3 \leq n \leq 16$ were reoptimised at the DFT level using the B3LYP exchange-correlation functional and 6-311++G** basis set. The results are shown in Figure 6.9. Each structure is accompanied by three labels; n, the number of water molecules, m, the energetic ordering of the rigid-body minimum from which it was re-optimised, and ΔU, the difference in energy between the minimum in question and the rigid-body global minimum having been reoptimised and evaluated at the DFT level.

Correlation plots of the energies of the reoptimised DFT minima against the original rigid-body minima for $3 \leq n \leq 16$ are presented in Figure 6.10. The degree of correlation varies with n, but in general, for the twenty lowest-energy rigid-body isomers considered, it is quite poor. The correlation is particularly poor for cases where the oxygen-thiocyanate skeleton of the rigid-body and DFT global minima differ (such as $n = 11$ and 12). An exception to this is the $n = 16$ system, which has relatively good correlation between the empirical and DFT minima, and is probably due to all twenty of the lowest energy rigid-body isomers sharing the same oxygen-thiocyanate skeleton. The empirical potential seems to be underestimating the difference in energy between different isomers when compared to the re-optimised DFT calculations. This is particularly prevalent for $n = 11$ and 12, where the energy range of the

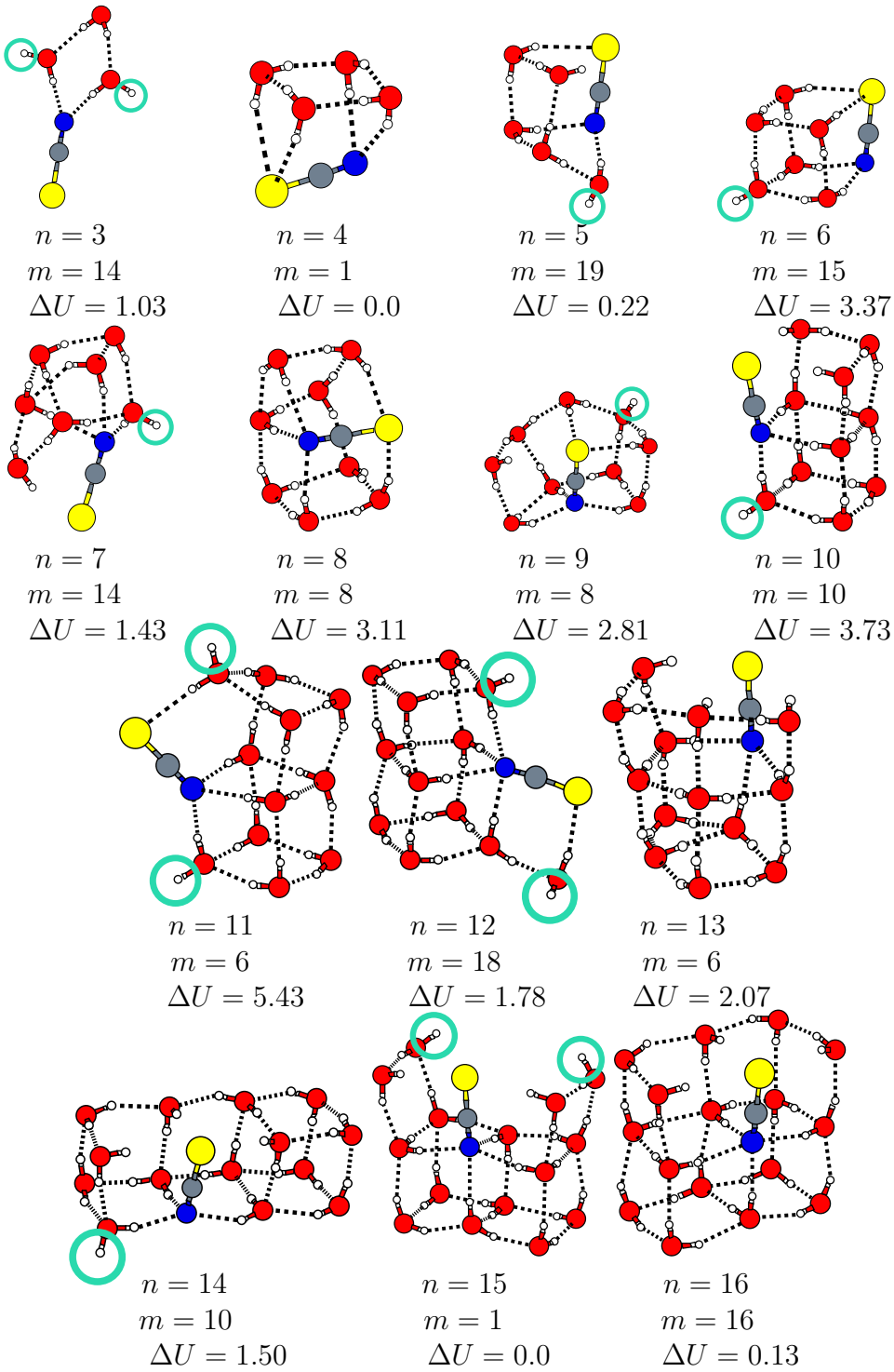


Figure 6.9: Structures of the putative global minima for $SCN^-(H_2O)_n$ clusters where $3 \leq n \leq 16$, computed at the DFT level using the B3LYP exchange- correlation functional and 6-311++G** basis set. The structures were determined by re-minimising the lowest twenty structures at each n. The m signifies the energetic ordering of the rigid-body minimum from which the structure was minimised, and ΔU is the difference in energy between the minimum in question and the rigid-body global minimum having been reoptimised and evaluated at the DFT level. The semi-dangling OH bonds in n = 3, 5, 6, 7, 9, 10, 11, 12, 14 and 15 are highlighted. Energies are in kcal mol^{-1}.

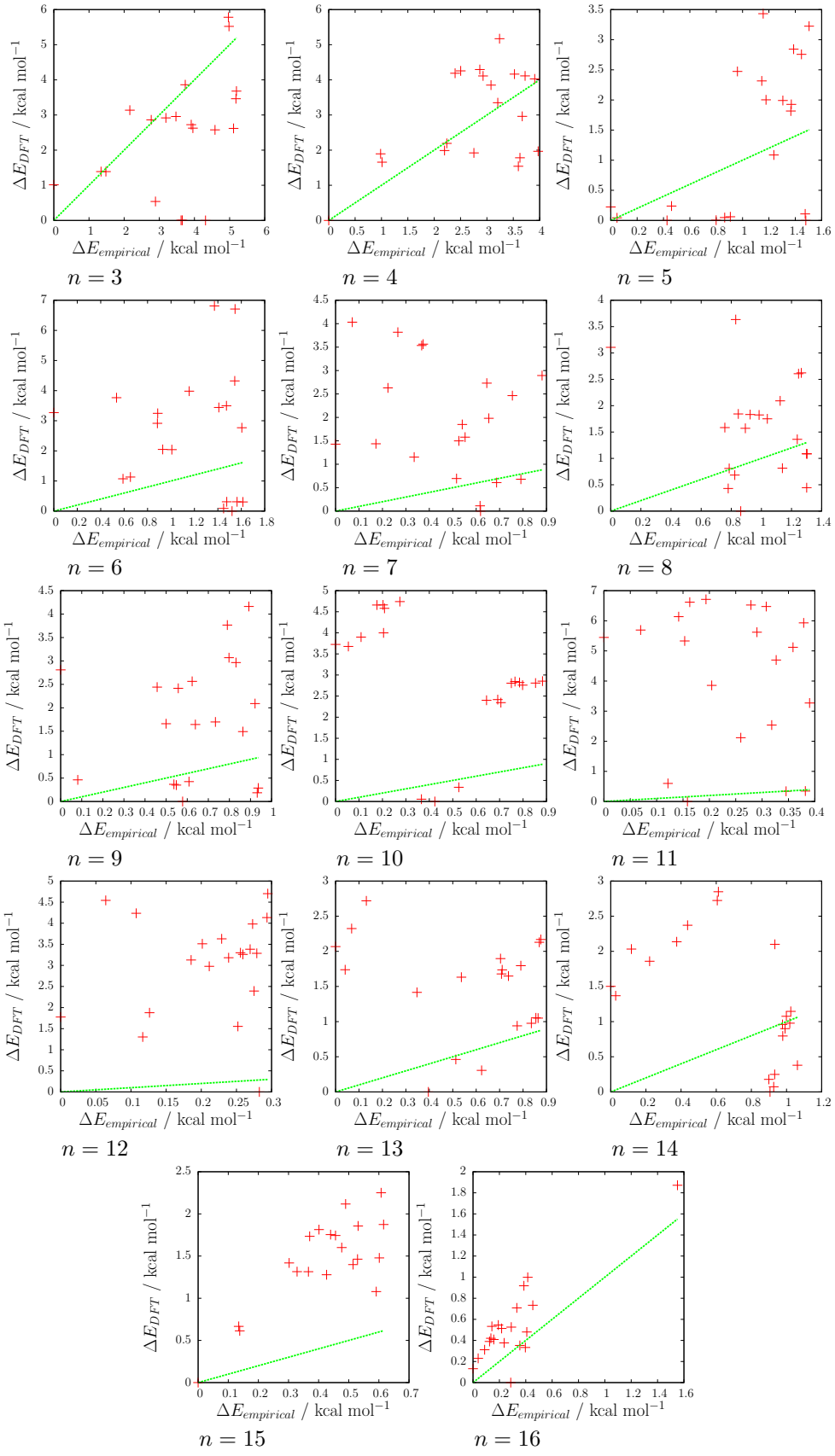


Figure 6.10: Correlation plot of re-optimised DFT minima computed at the B3LYP/6-311++G** level of theory against the original rigid-body minima, for $SCN^-(H_2O)_n$, with $3 \leq n \leq 16$.

lowest energy isomers at the two levels of theory differs by an order of magnitude. In spite of the poor correlation between the empirical and re-optimised DFT minima, the difference between the energy of the global minima at the two levels of theory is comparatively small, with the majority differing by less than 2.0 kcal mol^{-1}.

Structural analysis of the re-optimised global minima reveals one general trend: the empirical potential over-estimates the strength of the water-thiocyanate interaction, particularly with the sulfur atom. This is evidenced by three factors in the DFT calculations; the increased protrusion of the sulfur atom from the body of the cluster, the relocation of the thiocyanate ion to lower co-ordinated sites within the clusters, and the higher frequency at which water-thiocyante cubes are observed. At the DFT re-optimised level, water-thiocyanate cubes begin to appear at smaller n (six as opposed to ten water molecules) and do so more frequently. At $n = 6, 11, 12$ and 14, the number of water cubes or water-thiocyanate cubes increases, which has the double benefit of increasing the number of energetically stabilising structural sub-units, whilst also reducing the co-ordination number of the thiocyanate ion. Water cubes are so stabilising that the $n = 11$ and 12 clusters can essentially eject the ion from the structure and still remain the most energetically favourable geometry. The appearance of a single cube and one-dimensionally stacked fused cubes are reminiscent of pure water clusters, where these motifs are observed in the global minimum structures. Though the ion tends to move to lower co-ordinated sites, the mean co-ordination number of the nitrogen atom is still greater than that of the sulfur. The global minimum structures of the sixteen water molecule cluster share the same oxygen-thiocyanate skeleton as the empirical global minimum, but with a different hydrogen bonding network (they differ by two cycle inversions). Due to the large number of low-lying minima which share the same oxygen-thiocyanate skeleton, none of the structures which had been re-optimised had a different cage structure. Further investigation would be needed to determine if a more favourable heavy atom arrangement exists, though, given the propensity of the DFT calculations to favour water cubes, it is probably reasonable to assume that the skeleton is a good candidate structure (though a better hydrogen bond network may exist).

6.2 Structures and Landscapes of Hydrated Ferrocyanide Ions

IRPD spectroscopy indicates that the hydrated ferricyanide ion, ($Fe(CN)_6^{3-}$) requires at least eight water molecules for it to be electronically stabilised [130]. Though it is probable that the ferrocyanide ion, ($Fe(CN)_6^{4-}$), will require an even larger number of water molecules for it to be electronically stable, we present our global minimum results starting at eight water molecules. For clusters with $n \leq 15$ and $n = 17$, all eight basin-hopping runs located the same lowest common energy minimum. For $n = 15$ and 18, six and seven of the eight basin-hopping runs, respectively, located the same lowest common energy minimum. For $n = 19$ to 22, between two and three of the eight basin-hopping runs located a common lowest energy minimum. For $n \geq 23$, no two of the eight basin-hopping runs located a common lowest-energy structure. The structures of the global minima for $8 \leq n \leq 24$ are presented in Figure 6.11. To the best of our knowledge, putative global minima for hydrated ferrocyanide clusters have not been reported in the literature, so we have no previous results to compare with.

Central difference analysis of the global minimum structures (Figure 6.12) indicates that magic number clusters exist at $n = 24$ water molecules (and to a lesser extent 18 water molecules), and the lowest energy structures are shown in Figure 6.13. Peaks in the central difference are also observed at 35 and 41 water molecules, but visual inspection does not reveal any obvious structural features which might contribute to their comparative energetic stability, suggesting that the result is due to incomplete sampling at these larger sizes. The point group symmetries of the oxygen skeletons of the $n = 18$ and 24 global minima are C_2 and D_{4h}, respectively (see Figure 6.13 for further illustrations). For all sizes studied, the ferrocyanide ion is fully solvated, sitting centrally in the cluster. The geometries of the hydrated ferrocyanide global minima follow clear trends, with the water molecules assembling into structural sub-units. With a few minor exceptions, the global minimum water structure for all $n \leq 24$ are composed exclusively of said sub-units. In this thesis, the sub-units are identified as being either edge units, *closed* trimer units, *open* trimer units or terminal units

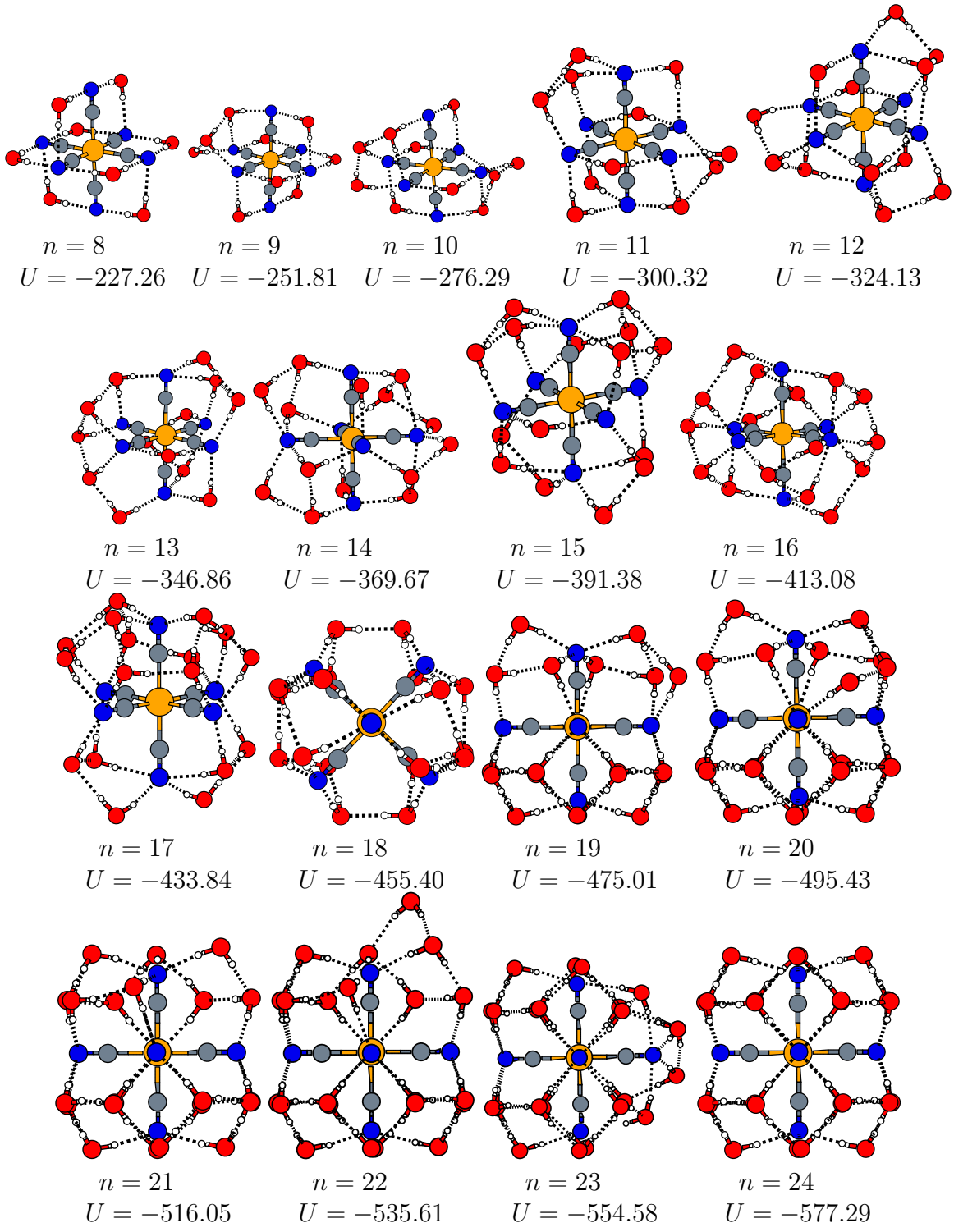


Figure 6.11: Structures and binding energies, U, of the putative global minima for $Fe(CN)_6^{4-}(H_2O)_n$ clusters where $8 \leq n \leq 24$. Energies are in kcal mol^{-1}.

(see Figure 6.14). An edge unit is composed of a single water molecule which forms hydrogen bonds with two nitrogen atoms, along the edge of the octahedron (see the global minimum structure for $n = 8$ in Figure 6.11). Open and closed trimeric units are each composed of

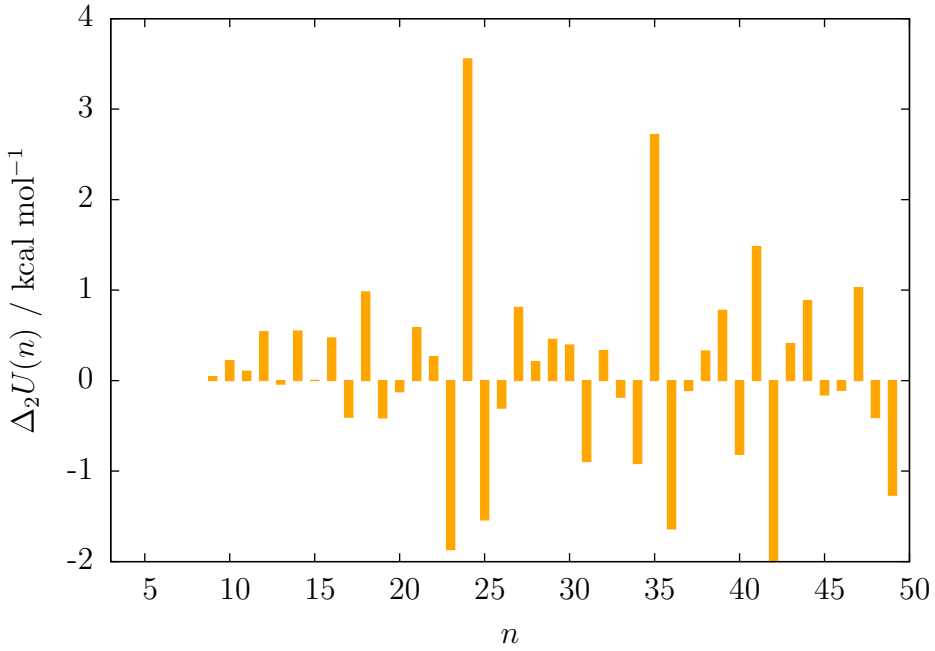


Figure 6.12: Central difference approximation to the second derivative of the energy of the putative hydrated ferrocyanide global minima, $\Delta_2 U(n)$, as a function of n.

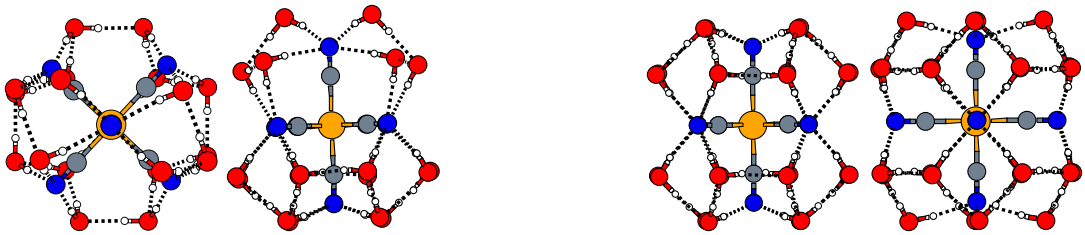


Figure 6.13: Structures of the observed 18 (left) and 24 (right) water molecule magic number clusters of the hydrated ferrocyanide system. Two perspectives are shown for each cluster to aid visualisation.

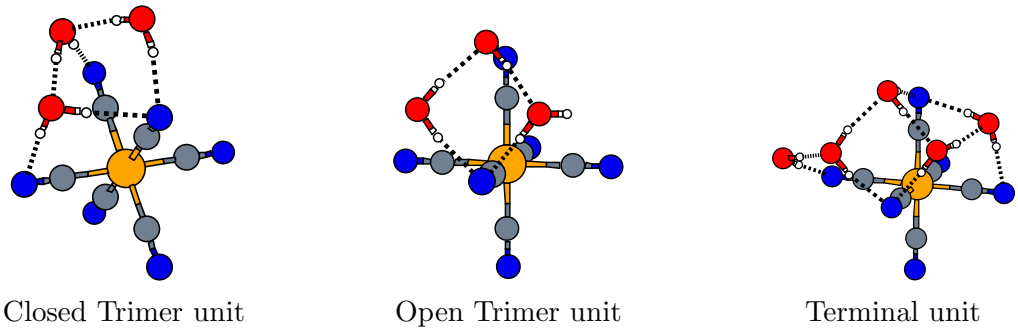


Figure 6.14: Common structural motifs for the $Fe(CN)_6^{4-}(H_2O)_n$ clusters for $8 \leq n \leq 24$.

three water molecules, making them the largest sub-units in the system upto $n = 24$. Closed trimer units are superficially similar to the trimer rings observed in hydrated sulfate clusters (see Figure 5.11 in Chapter 5) but do not form a closed cycle of hydrogen bonds, instead having one of the water molecules edge bond with the ferrocyanide. Open trimer units also

do not form a closed cycle of hydrogen bonds, but replace the edge bonding water molecule with a water molecule which exhibits a dangling OH bond. Open trimeric units are never observed in isolation, and are either adjacent to another open unit or encapsulated by a terminal unit. Terminal units are composed of two water molecules which sit either side of one or more open trimers. Half the terminal unit edge bonds with the ferrocyanide and accepts the dangling OH bond of the open trimer unit, whilst the other half donates a hydrogen bond to the opposite end of the open trimer. The terminal unit illustration in Figure 6.14 portrays the terminal unit flanking a single open trimer, though terminal structures which sandwich two or three open trimer units are also observed (the global minimum structure for $n = 17$ exhibits a terminal unit encapsulating three open trimers). A final comment on the structure of the sub-units: the edge and closed trimer units are both discrete structures, whereas the open trimer and terminal units are continuous and most be adjoined either to themselves or to each other. An (almost) complete representation of the observed global minimum structures can be gained be counting the number of sub-units in each structure. The frequency with which each sub-unit appears in the global minima geometry is reported in Table 6.2.

n	Edge	Closed Trimer	Open Trimer	Terminal
8	8	0	0	0
9	6	1	0	0
10	4	2	0	0
11	5	2	0	0
12	3	3	0	0
13	2	2	1	1
14	2	4	0	0
15	1	3	1	1
16	0	2	2	2
17	0	2	3	1
18	0	2	4	0
19	0	1	5	1/2
20	0	1	5	1
21*	0	0	6	1
22*	0	0	6	1
23*	0	0	6	1
24	0	0	8	0

Table 6.2: Number of labelled sub-units in hydrated ferrocyanide global minimum geometries as a function of the number of water molecules, n, for $8 \leq n \leq 24$. Clusters marked with * exhibit structural features not fully described by the labelled sub-units.

For small cluster sizes ($8 \leq n \leq 9$), the majority of water molecules are in edge units, interacting almost solely with the ion. The number of edge units decreases as the number of water molecules increases, and they cease to be observed for $n \geq 16$. In the mid- size range ($10 \leq n \leq 15$), most water molecules form closed trimer sub-units. Closed units persist until $n \geq 20$, at which point they are subsumed by open trimer units. Open trimer units first appear at $n = 13$, and become the dominant feature at larger sizes ($n \geq 17$). For $n \geq 16$ (coinciding with the disappearance of edge units), the water structure organises into two distinct hemispheres with a principal axis coincident to the four-fold symmetric corner-to-corner axis of the ferrocyanide octahedron. Between the two hemispheres there are no water-water hydrogen bonds. As the number of water molecules increases, the open trimers join to form larger continuous units. For global minimum structures containing 18 or more water molecules, four open trimer units join to form a single extended structure, which completely spans one hemisphere of the cluster. This behaviour culminates in the global minimum geometry of the $n = 24$ cluster, which exhibits two hemispheres of four open units. The stabilizing effect of the hemispheres is such that central difference analysis indicates that the $n = 24$ global minimum is the most stable structure relative to its neighbours by a considerable margin. There are a few exceptions for which a decomposition into sub-units does not give a complete description of the structure. The $n = 19$ global minimum contains half a terminal unit, with only one end of a lone open trimer unit being capped. The structures $21 \leq n \leq 23$ (marked with a * in Table 6.2) each contain one, two and three water molecules, respectively, which do not participate in one of the identified sub-units. All of the structures observe the same basic template: one hemisphere composed of four open units and the other hemisphere containing two open trimers flanked by a terminal unit, which accounts for 20 water molecules. The additional water molecules are then positioned elsewhere in the cluster. The $n = 23$ global minimum is the only cluster containing more than 15 water molecules which does not have two distinct hemispheres, with the three additional water molecules forming a bridge between the two.

A databases of minima and transition states for the $n = 8$ cluster containing 10,758 minima and 72,338 transition states has been compiled. Figures 6.15 and 6.16 show the dis-

connectivity graph and three low-energy isomers of $Fe(CN)_6^{4-}(H_2O)_8$, respectively. The

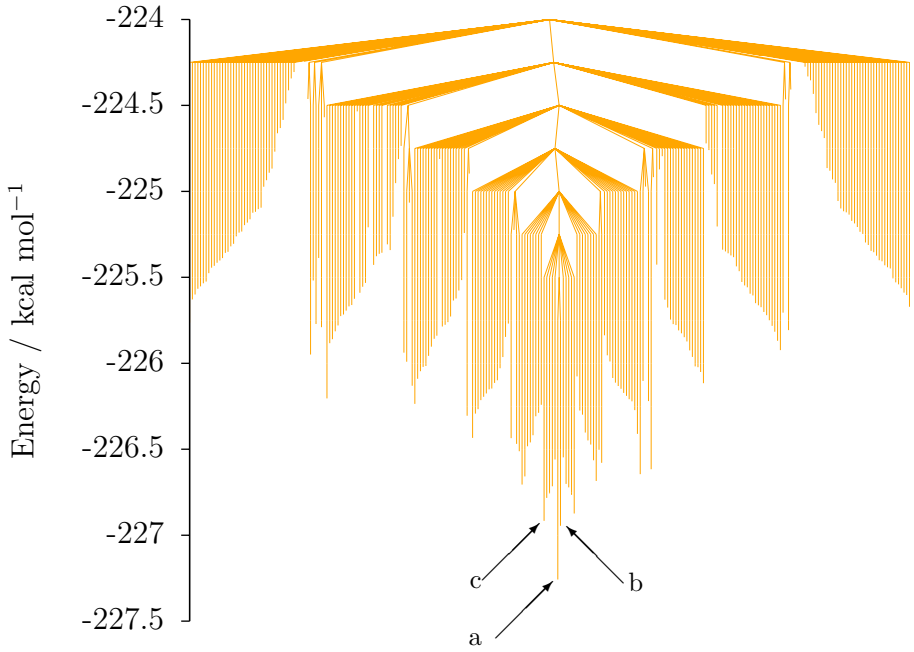


Figure 6.15: Disconnectivity graph for $Fe(CN)_6^{4-}(H_2O)_8$ containing 263 minima and 488 transition states connected to the global minimum below -224.0 kcal mol^{-1}.

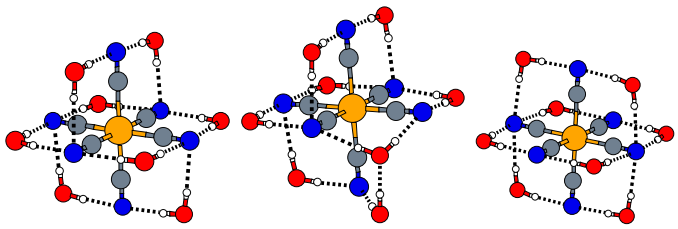


Figure 6.16: The three lowest-energy isomers of $Fe(CN)_6^{4-}(H_2O)_8$ (a-c), as labelled on the disconnectivity graph in Figure 6.15. ΔU is the energy of a given isomer above the global minimum. Energies are in kcal mol^{-1}.

$Fe(CN)_6^{4-}(H_2O)_8$ energy landscape is reasonably funnelled, with an energetically distinct global minimum and asymmetric barriers connecting the global minimum to energetically less favourable isomers. The lack of water-water interactions in the cluster means that unlike in other hydrated ion systems, isomers which share an oxygen-ion skeleton but differ in the hydrogen bond network are not observed. As with the hydrated sulfate systems, the landscape of the eight water molecule cluster suggests that locating the optimal oxygen skeleton is key to determining the global minimum structure.

A final comment on the optimisation of the ferrocyanide clusters; the number of low energy minima which posses enough water-water hydrogen bonds to support a hydrogen bond cycle is so low that the cycle inversion moves employed for the other hydrated clusters proved ineffective. To improve the efficiency of the searching for these systems, cycle inversion moves were not utilised. Instead they were performed subsequently on low energy structures which supported cycles to investigate their cycle inversion isomers.

6.3 Structures and Landscapes of Hydrated Ferricyanide Ions

The research presented in this Section was partially conducted by Deepak Sharma, a summer research student in the Johnston group. Deepak was responsible for performing global optimisation searches for clusters containing more than 18 water molecules, and for building the database of minima and transition states for the eight water molecule hydrated cluster. IRPD spectroscopy indicates that the hydrated ferricyanide ion requires at least eight water molecules for it to be electronically stabilised [130]. For clusters with $n \leq 12$ and $17 \leq n \leq 18$, all eight basin-hopping runs located the same lowest common energy minimum. For $13 \leq n \leq 16$ fewer than eight of the independent searches found a common lowest energy minimum, with the putative global minimum for $n = 15$ observed only once. For $n \geq 19$, no two of the eight basin-hopping runs locate a common lowest-energy structure. Compared to the other hydrated ion systems studied, the search statistics for the hydrated ferricyanide ion are atypical, and a discussion and comparison will follow in Section 6.5. The structures of the global minima for $8 \leq n \leq 18$ are presented in Figure 6.17. To the best of our knowledge, putative global minima for hydrated ferricyanide clusters have been reported for the $n = 8$ cluster only, so we have no previous results to compare with for larger sizes. For the eight water molecule clusters, a global minimum and low energy isomer structures calculated at the B3LYP/LACVP++** level of theory by DiTucci *et al.* [130]. Although the structures were generated by hand and not via a global optimisation routine, they can still provide a useful benchmark. However, the literature results predict that the global mini-

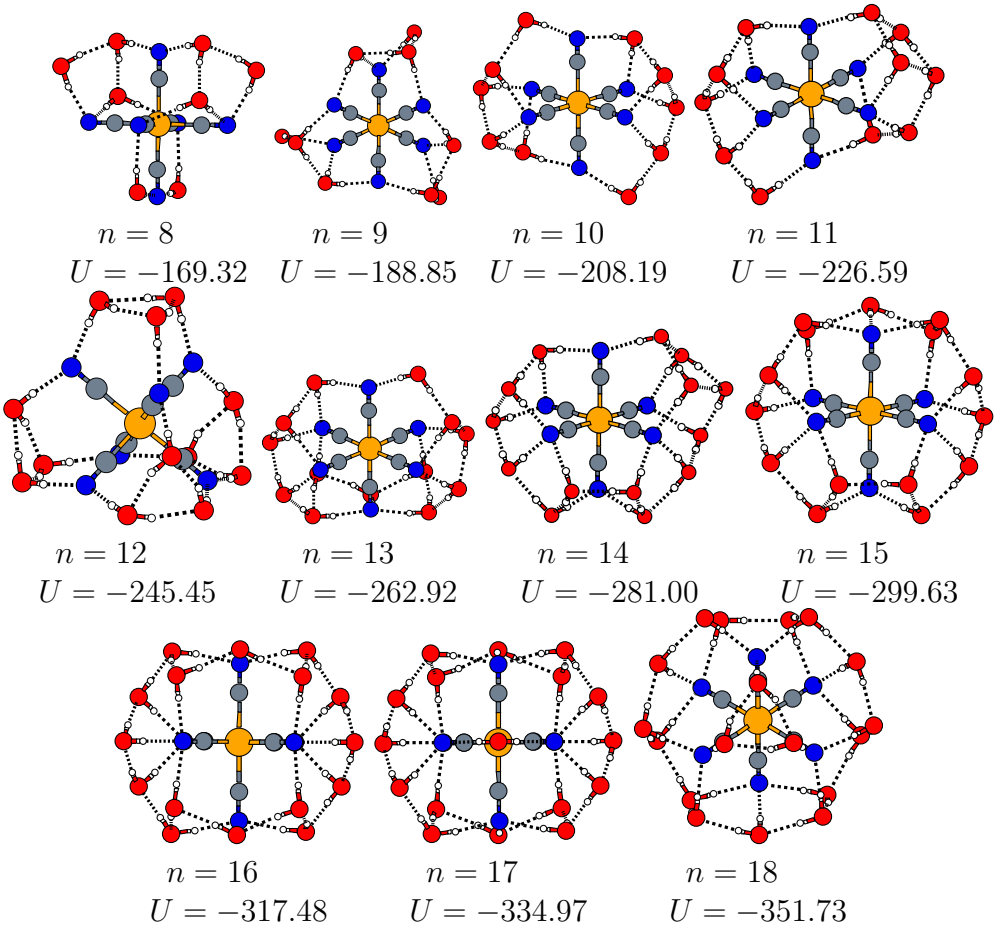


Figure 6.17: Structures and binding energies, U, of the putative global minima for $Fe(CN)_6^{3-}(H_2O)_n$ clusters where $8 \leq n \leq 18$. Energies are in kcal mol^{-1}.

mum structure should contain no edge units, which is inconsistent with the results at the empirical level for $n = 8$. Results from re-optimisation of these empirical isomers tentatively suggest that they maybe more stable, though these results are still being performed and are not included in this thesis.

Central difference analysis of the global minimum structures (Figure 6.18) indicates that no magic number clusters exist at below 18 water molecules (the limit of reliable global minimum sampling). Peaks in the central difference are observed at $n = 26$, 27, 42, 47 and 49, though visual inspection reveals no obvious structural features which might confer stability to these geometries, which suggests that the results maybe a consequence of poor sampling at larger sizes.

As with the ferrocyanide ion, the ferricyanide ion is fully solvated, sitting centrally in the cluster for all sizes studied. The geometries of the hydrated ferricyanide global minima follow

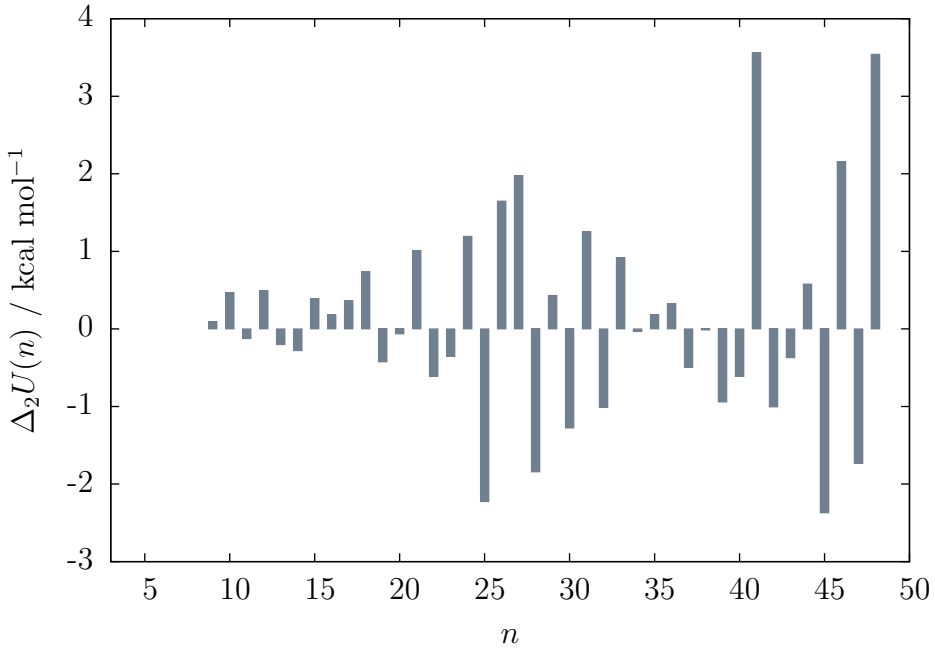


Figure 6.18: Central difference approximation to the second derivative of the energy of the putative hydrated ferricyanide global minima, $\Delta_2 U(n)$, as a function of n.

similar trends to the ferrocyanide, though not so rigidly and with more exceptions. Unlike in the hydrated ferrocyanide system, the water molecules are organised about the three-fold axis of the ferricyanide ion (in $Fe(CN)_6^{4-}(H_2O)_n$, the water molecules are broadly organised about the corner-to-corner axis of the ferrocyanide ion). Similar structural sub-units are observed in both systems, but do not form as complete as a description of the hydrated 3- ion as they do for the 4- ion, and in some instances the sub-units are not present at all.

Global minimum structures containing eight, nine and ten water molecules follow the same structural progression as the ferrocyanide cluster and are composed entirely of edge, open trimer, closed trimer and terminal sub-units. Trimeric water rings equivalent to those reported for hydrated sulfate clusters (see Figure 5.11 in Chapter 5) are observed in the global minimum structures for $n = 12$ and 18 (the $n = 12$ ferricyanide and sulfate global minima are exact analogues of one another, and both systems exhibit the same cycle inversion isomers). As reported previously, the search success rate falls from agreement between all eight searches at $n = 12$ to a single search locating the putative global minimum at $n = 15$. Normally, such poor concurrence between searches would imply that the landscape is difficult to search (for a given search method), and the lowest energy minimum found would not

be a good candidate global minimum. However, given that the putative global minimum structure for $n = 15$ is approximately 1.00 kcal mol^{-1} more stable than the next isomer and that the oxygen skeleton is C_{3v} symmetric (the second isomer is C_1), it is reasonable to assume that the reported $n = 15$ structure is a good candidate global minimum (in spite of the discouraging search statistics). A brief discussion as to why this global minimum may be difficult to locate (and why others are comparatively much easier) is given in Section 6.5. The water molecules in the $n = 15$ global minimum span the equator of the ferricyanide ion, perpendicular to the three-fold axis of the ion, and form a single connected structure. A similar equatorial structure is also present in the $n = 16$ global minimum, which has C_{2v} point group symmetry. The $n = 17$ structure is identical to the $n = 16$ global minimum, with the extra water molecule occupying an edge bonding site. The $n = 18$ structure is identical to the $n = 15$ structure, but with the extra three water molecules participating in a trimeric ring. The 17 and 18 water molecule global minima have C_{2v} and C_{3v} point group symmetries, respectively.

A databases of minima and transition states for the $n = 8$ cluster containing 13,279 minima and 131,126 transition states has been compiled. Figures 6.19 and 6.20 show the disconnectivity graph and three low-energy isomers of $Fe(CN)_6^{3-}(H_2O)_8$, respectively.

The $Fe(CN)_6^{3-}(H_2O)_8$ landscape is comparable to the $Fe(CN)_6^{4-}(H_2O)_8$ landscape, in that it features a number of asymmetric barriers connecting the global minimum to higher energy isomers, resulting in a landscape which exhibits some funnelling character. However, the ferricyanide system contains more frustration, with larger barriers separating minima, and a number of isomers with energies similar to the global minimum energy. In this regard it is more similar to the other hydrated ion systems (sulfate, thiocyanate and perchlorate), which have a greater degree of frustration and due to the degeneracy of the oxygen skeleton. The two isomers of $Fe(CN)_6^{3-}(H_2O)_8$ labelled b and c in Figure 6.20 share the same structural sub-unit signature as the global minimum (two edges, two closed trimers) but differ in either their relative positioning about the ion (isomer b) or in the relative directionality of the hydrogen bonds in the closed trimers (isomer c). For the eight water molecule hydrated ferricyanide cluster, finding a good oxygen skeleton is not critical to determining

the global minimum, as a number of energetically competitive structures exist with different skeletons. As with the ferrocyanide ion, cycle inversion moves proved ineffective due to the lack of closed hydrogen bond cycles in the low energy structures. Instead, the searches were conducted without hydrogen bond optimisation moves, and the cycle inversion moves were then performed retroactively on promising candidate isomers.

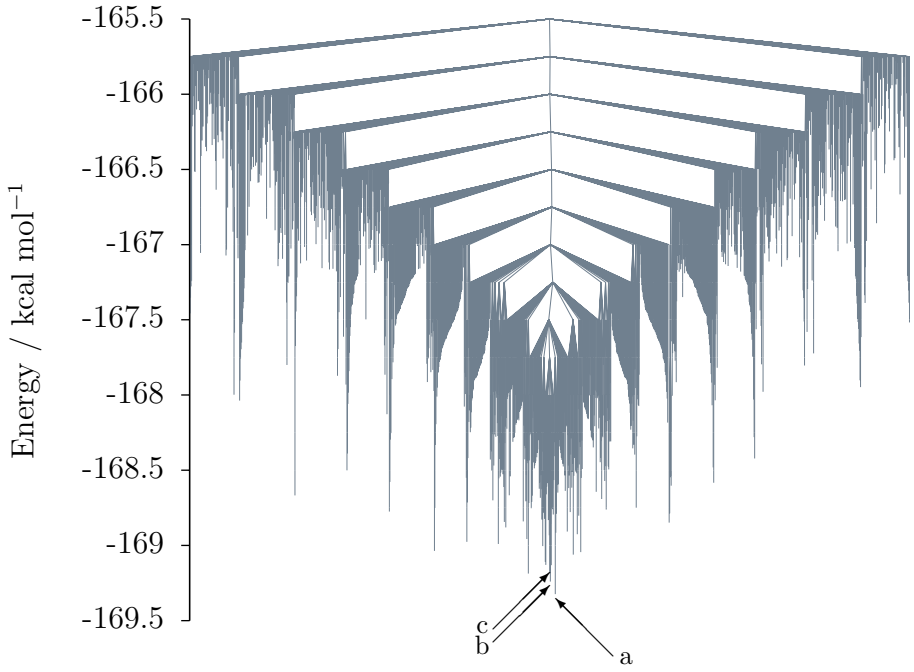


Figure 6.19: Disconnectivity graph for $Fe(CN)_6^{3-}(H_2O)_8$ containing 1442 minima and 3791 transition states connected to the global minimum below -165.5 kcal mol^{-1}.

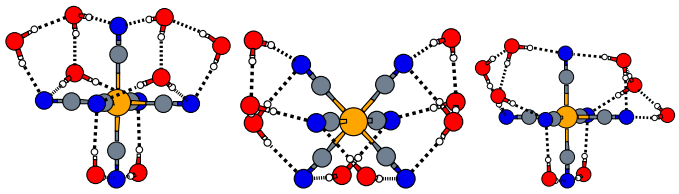


Figure 6.20: The $Fe(CN)_6^{3-}(H_2O)_8$ global minimum (a) and the 2^{nd} and 4^{th} lowest-energy isomers (b-c) respectively, as labelled on the disconnectivity graph in Figure 6.15. ΔU is the energy of a given isomer above the global minimum. Energies are in kcal mol^{-1}.

6.4 Structures and Landscapes of Hydrated Perchlorate Ions

The data presented in this Section were collected by John Hey, and forms part of his Masters' dissertation. It is included here for completeness and for comparison with other hydrated ion systems. For clusters with $n \leq 15$, all eight basin-hopping runs located the same lowest common energy minimum. For $n \geq 16$, no two of the eight basin-hopping runs locate a common lowest-energy structure, and thus the putative 'global minima' presented at this size have been observed by only one of the independent basin-hopping searches. Hence they may be representative low-lying minima, rather than the true global minimum. The structures of the global minima for $3 \leq n \leq 15$ are presented in Figure 6.21. The semi-dangling OH bonds in $n = 5$, 12, 13 and 15 are highlighted. Central difference analysis (Figure 6.22) of the global minimum energies suggests that our searching has not sampled any magic number structures. For small sizes (fewer than nine water molecules), an alternating even-odd pattern in stability is observed, similar to the behaviour of homogeneous TIP4P water [173], and corresponds to the appearance of one, two and three water squares. The maximally stable global minimum structure as determined by central difference analysis is the $n = 37$ structure. A visual inspection of the cluster geometry does not reveal any obvious structural traits which might confer greater stability, and so the result may be a consequence of the difficulty of sampling at larger system sizes.

As with the thiocyanate ion, the perchlorate ion prefers to sit on the surface of the cluster, either sitting at low co-ordination sites or not becoming fully solvated at all for all the sizes studied. Unlike the other hydrated ion systems studied in this thesis, no recurring structural motifs are observed (such as trimeric water rings in $SO_4^{2-}(H_2O)_n$), and the low energy minima are structurally irregular. This complicates the analysis, as few global trends or themes can be gleaned. Water squares are a recurring feature in the global minimum structures, appearing in every size above three water molecules. One can make some tentative comparison with water clusters, and note that some of the perchlorate structures are structural analogues to the water cluster global minima. In Figure 6.23 two structural comparisons

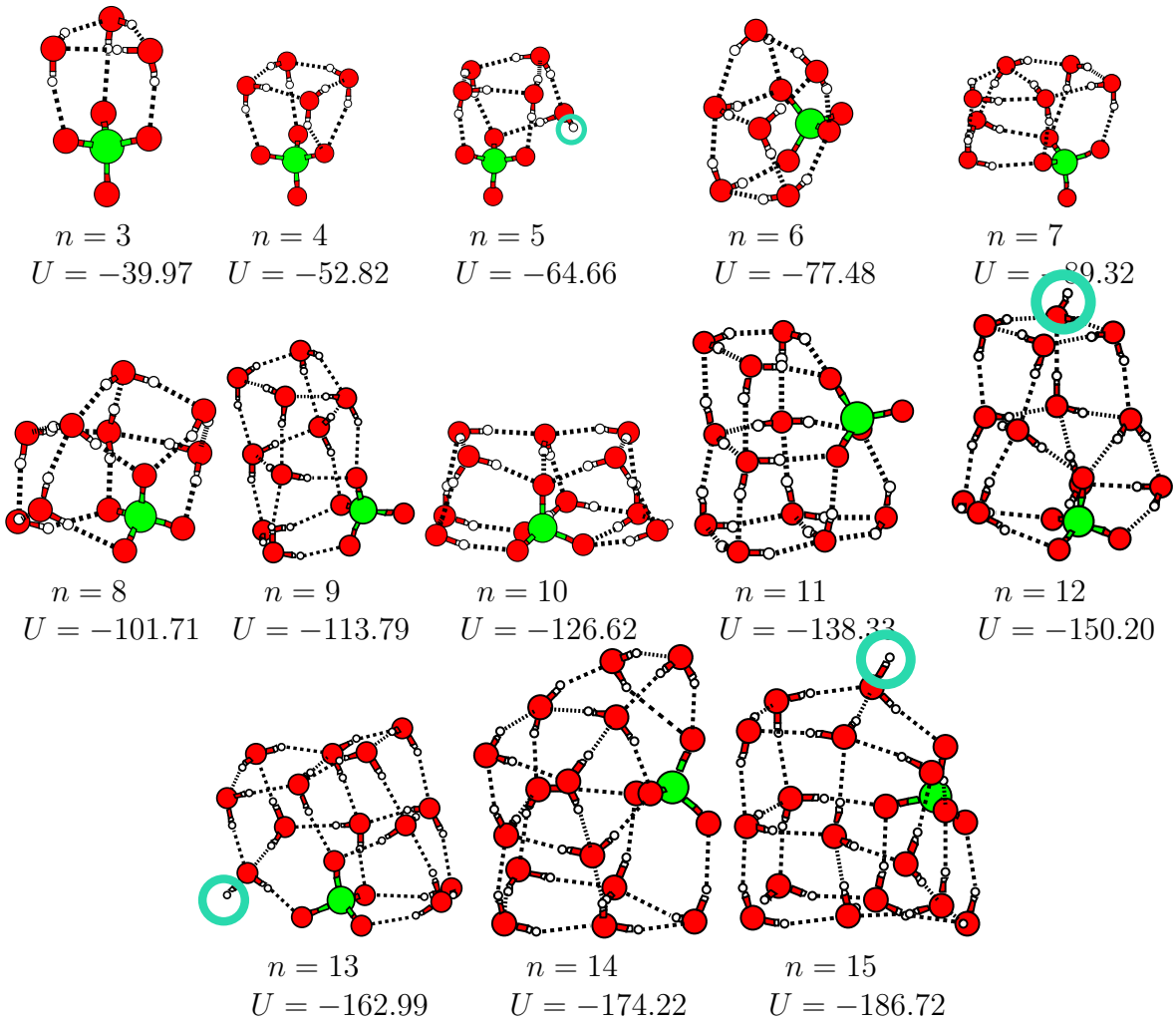


Figure 6.21: Structures and binding energies, U, of the putative global minima for $ClO_4^-(H_2O)_n$ clusters where $3 \leq n \leq 15$. The dangling OH bonds in $n = 5$, 12, 13 and 15 are highlighted. Energies are in kcal mol^{-1}.

are made; between $(H_2O)_{12}$ and $ClO_4^-(H_2O)_9$ and between $(H_2O)_{14}$ and $ClO_4^-(H_2O)_{12}$. The $ClO_4^-(H_2O)_9$ structure mimics the $(H_2O)_{12}$ global minimum, with the water squares preserved (or substituted for water-perchlorate squares) on three of the four long sides of the fused-cube. The $ClO_4^-(H_2O)_{12}$ preserves the same number of four and five membered water / water-perchlorate rings as the $(H_2O)_{14}$ global minimum, but with the relative positions of the rings permuted about the structure. Neither cases preserve the oxygen skeleton exactly, but they both seek to satisfy the same energetic requirements (to maximise four and five membered water rings) and to do so in the same manner.

Databases of minima and transition states for clusters in the size range $3 \leq n \leq 16$ have been compiled, and the results for $n = 12$ is presented: in this case, the database contains

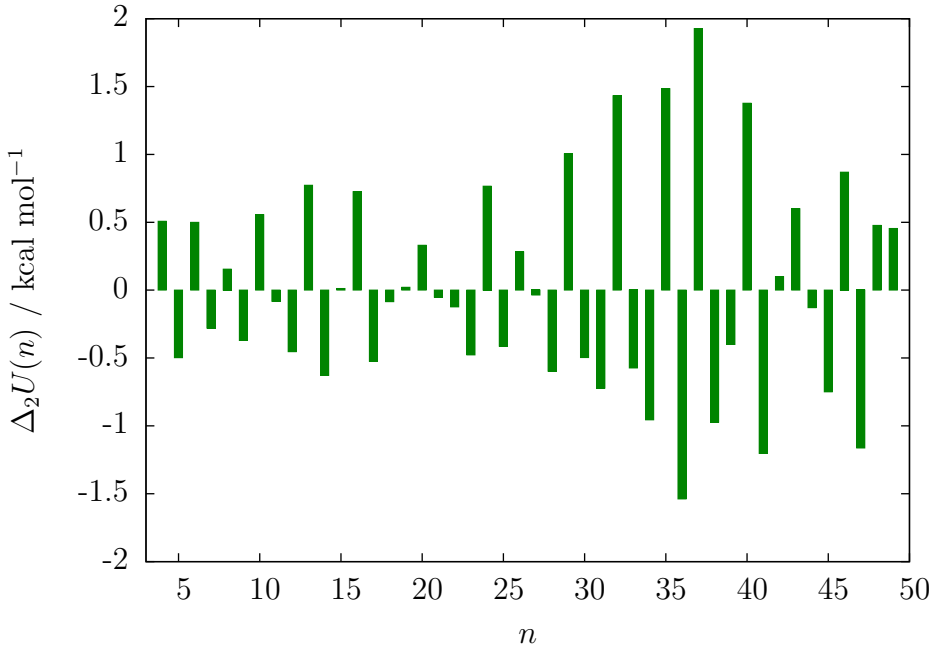


Figure 6.22: Central difference approximation to the second derivative of the energy of the putative hydrated perchlorate global minima, $\Delta_2 U(n)$, as a function of n.

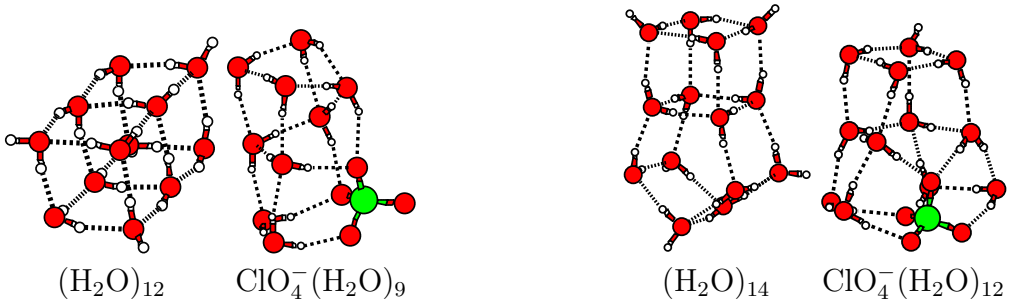


Figure 6.23: Structural comparison between hydrated perchlorate and pure TIP4P water global minima. TIP4P water structures are taken from the Cambridge Cluster Database [160].

13663 minima and 328656 transition states. The disconnectivity graph and seven lowest energy minima of $ClO_4^-(H_2O)_{12}$ are given in Figures 6.24 and 6.25, respectively.

Six of the seven lowest energy isomers share an oxygen skeleton, (differing from one another by cycle inversions of the hydrogen bond network) and are almost isoenergetic, (spanning an energy range of only 0.19 kcal mol^{-1}). Despite the similarities, the energetic barriers between the isomers are substantial (approximately 4.5 kcal mol^{-1}), leading to the landscape exhibiting a glassy topography.

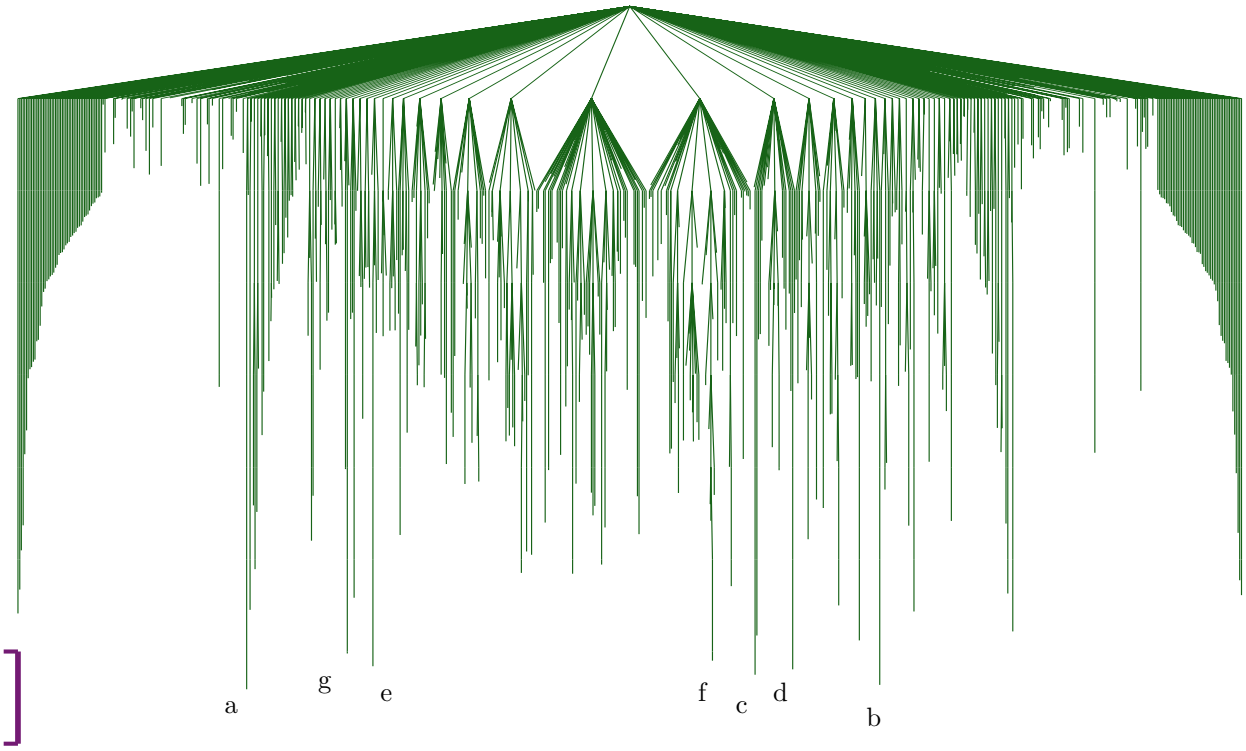


Figure 6.24: Disconnectivity graph for $ClO_4^-(H_2O)_{12}$ containing 718 minima and 840 transition states connected to the global minimum below -146.5 kcal mol^{-1}. The purple bar is a scale of 1 kcal mol^{-1}.

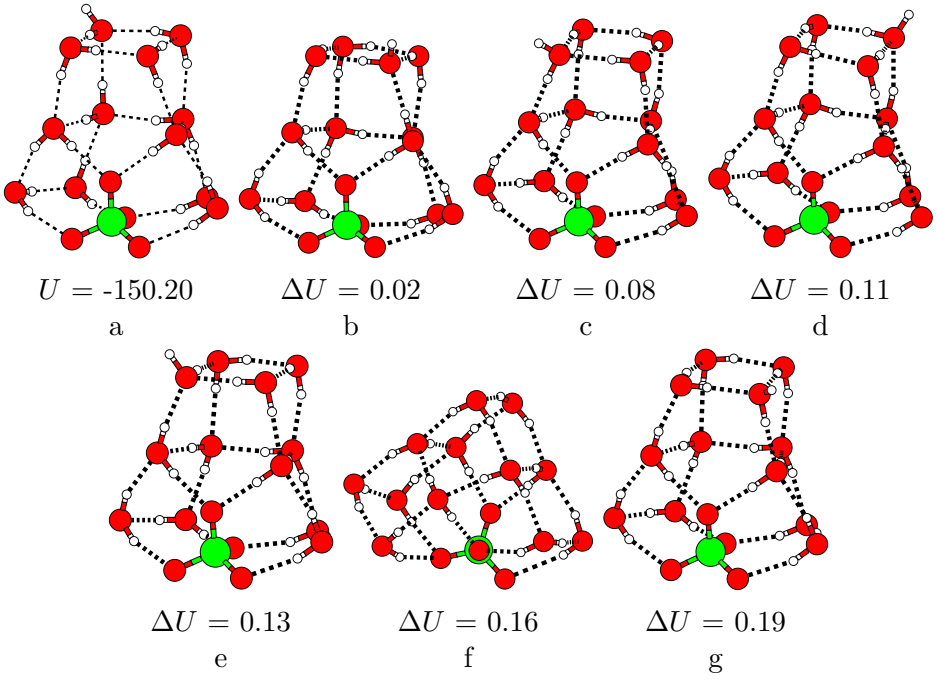


Figure 6.25: The $ClO_4^-(H_2O)_{12}$ global minimum (a) and the next six lowest-energy isomers (b-g), as labelled on the disconnectivity graph in Figure 6.24. ΔU is the energy of a given isomer above the global minimum. Energies are in kcal mol^{-1}.

6.5 Comparative Analysis of Hydrated Ion Geometries

Discussion and comparison of the five hydrated ion systems presented in this thesis is given below. Results discussed include the searchability of the different systems, their propensity

to exhibit dangling OH bonds and the hydrogen bond acceptor-to-donator ratio of water molecules in each system.

The number of independent basin-hopping searches which found the reported global minimum for each hydrated ion system is shown as a function of the number of water molecules in Figure 6.26.

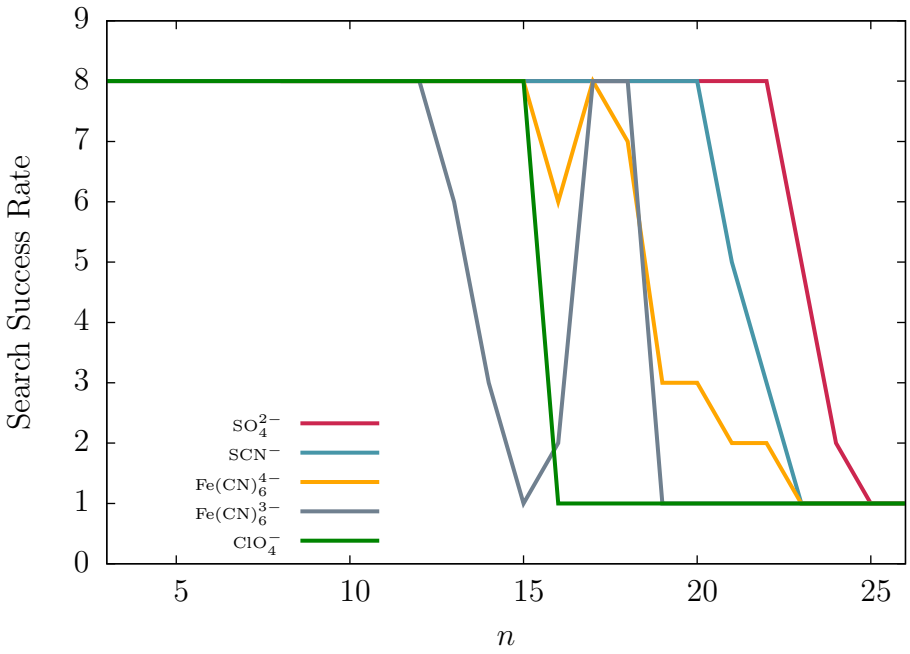


Figure 6.26: Number of independent basin-hopping runs which located a common lowest energy minimum as a function of the number of water molecules, n, for $3 \leq n \leq 26$. None of the systems studied exhibited a putative GM observation frequency greater than one for $n \geq 26$.

The number of independent searches which located the reported global minimum is interpreted as a measure of the ease with which a global minimum structure can be located for a given system. Of the five hydrated ion systems investigated, the sulfate ion is the most readily searchable, with concurrency between all searches maintained until $n = 23$, then dropping precipitously to zero concurrency at $n = 24$. The thiocyanate and ferrocyanide ions both exhibit similar searchability statistic to one another, with search concurrency trailing to zero at 23 water molecules each. The hardest landscapes to search are those of the perchlorate and ferricyanide ions, which lose concurrency at 19 and 16 water molecules, respectively (the ferricyanide ion is unique amongst the system studied in that searchability drops in the mid-size range before retuning to full concurrency at 17 water molecules, a feature which is discussed below).

Typically, one would expect hard-to-search systems to be those with a glassy landscape, trapping the Markov chain in one of a number of deep metastable funnels. Conversely, searchable systems might be expected to have funnelled landscapes. However the landscape analysis reported in the previous sections reveals no such correlation between search efficiency and landscape topography. The sulfate, thiocyanate and perchlorate ions all have landscapes exhibiting varying degrees of frustration, and yet they experience a marked difference in searchability (one should note that a large portion of this frustration is due to barriers between isomers with similar oxygen skeletons but differing hydrogen bond networks, which would be traversed by cycle inversion moves). On the other hand, the ferri-and-ferrocyanide ions both have comparatively funnelled landscapes, and yet give rise to very different search statistics.

Though it is not entirely clear why one system should be more searchable than another, the manner in which the water molecules arrange themselves structurally in different systems could provide an explanation. Ions which encourage the water molecules into small discrete or modular sub-units from which larger structures are constructed are the least challenging to search, especially if the sub-units are themselves energetically stable. Each of the three searchable ion systems achieve this by either encouraging 'natural' water structures such as cubes and pentagonal prisms (thiocyanate), or forcing the water to adopt small discrete units such as trimers and hexagons (sulfate and ferrocyanide). To illustrate why this behaviour might expedite global optimisation, consider how a basin-hopping process searches the $SO_4^{2-}(H_2O)_{12}$ landscape, where the global minimum has all 12 water molecules engaged in a one of four trimeric rings. Because each trimer ring is itself the most energetically stable way to arrange three water molecules, when the Markov chain encounters one, it is unlikely to be subsequently dismantled. This allows the search process to locate trimer rings sequentially, with each new ring further stabilising the structure until all four have been found. In other instances, large connected water structures such as those observed for $SCN^-(H_2O)_{16}$ can be easily found because the structure is robust to errors or defects. Once a suitable (though not necessarily optimal) oxygen skeleton is determined, it is sufficiently stable that the defects can be subsequently 'fixed' before the structure is disassembled. There is also

very little compromise required in their low energy geometries. We define compromise to indicate a structure composed of sub-units which themselves are *not* the most energetically stable sub-unit. A structure containing a lot of compromise is one in which energetically sub-optimal sub-units combine to form an energetically optimal whole structure. Conversely, optimal structures composed of optimal sub-units contain no compromise. Low compromise structures feature heavily in the global minima of the searchable ions, i.e. the $SCN^-(H_2O)_{16}$ global minimum is basically the $SCN^-(H_2O)_{10}$ global minimum with added cubic units.

Neither the ferricyanide nor perchlorate systems have global minimum structures which contain a large number of energetically favourable discrete or modular sub-units, and both are difficult to search.

The ferricyanide ion global minima form geometries composed of a single connected water structure, but unlike the hydrated thiocyanate global minima they are not robust to errors. A good example is the 15 water molecule ferricyanide global minimum, which was only located by a single independent search, suggesting that it is a particularly challenging minimum to detect. The water structure is a single connected unit which is not robust to defects (the next lowest energy isomers are approximately 1 kcal mol^{-1} lower in energy and with a markedly different water structure). Due to the lack of robustness in the structure, a search process could not sequentially build the observed geometry and must instead locate the whole structure in a single step. Other ferricyanide global minima geometries exhibit a larger degree of compromise than the more searchable ions. For example, the $Fe(CN)_6^{3-}(H_2O)_{12}$ global minimum structure has all twelve water molecules engaged in one of four trimeric water rings (analogous to the $SO_4^{2-}(H_2O)_{12}$ global minimum) even though in isolation a trimeric water ring is approximately 0.7 kcal mol^{-1} less stable than a closed trimer unit, which is far more commonly observed in the ferri/ferrocyanide systems. A final note on ferricyanide searchability: despite the difficulty faced when trying to determine the 15 water molecule global minimum structure, the search statistics improve at larger sizes, returning to full concurrency at $n = 17$ and 18. This is curious given that the $n = 18$ structure is simply the $n = 15$ global minimum structure, with a trimer ring positioned at one of the open surfaces of the ion. In this case, the trimer ring must help to stabilise the

incomplete $n = 15$ structure in some manner, though how (if indeed at all) remains an open question.

The hydrated perchlorate clusters are even less searchable than their ferricyanide counterparts, though for slightly different reasons. The hydrated perchlorate ion does not form modular structures like the searchable ions, nor does it form fragile structures like the ferricyanide. Instead it occupies a grey area, where it is not strongly interacting enough to pattern the waters into non-water like structures like the sulfate and ferri/ferrocyanide ions, but it is too intrusive and geometrically dissimilar to the water molecules to encourage water structure to form like the thiocyanate. Instead, the structures are highly compromised, with water molecules forced away from energetically stable water-like structural sub-units by the perchlorate, but without any great energetic reward for doing so. The result is that the search process is retarded, as the global minimum structures become highly specific to a given number of water molecules.

In Figure 6.27 the Boltzmann-weighted mean number of dangling OH bonds, $\bar{f}$, is plotted as a function of the number of water molecules for each of the five hydrated ion systems (see Equation 5.2 in Chapter 5). A temperature weighting of $T = 130$ K is used to be consistent with experiment (where appropriate) [141, 130], and it is assumed that at this temperature the free energy can be estimated from the potential energy alone (i.e. that the entropic contribution to the free energy from each isomer is approximately the same). An examination of Figure 6.27 reveals two behaviour regimes; one in which the number of dangling OH bonds are completely suppressed up to a critical value of n (which may be larger than the size range studied), and another in which dangling OH bonds begin to appear at low n, and grow steadily with increasing system size. The sulfate, ferricyanide and ferrocyanide ions all belong to the first regime. The sulfate results are discussed in depth in Chapter 5, where it was determined that the ion suppressed the appearance of dangling OH bonds until approximately 43 water molecules, in agreement with experimental data [141]. For the ferricyanide ion, the water structure experiences a suppression of dangling OH bonds until at least 50 water molecules, provisionally in agreement with experimental IRPD spectra [130]. Experimental data suggest that free OH bonds begin to appear in hydrated ferricyanide

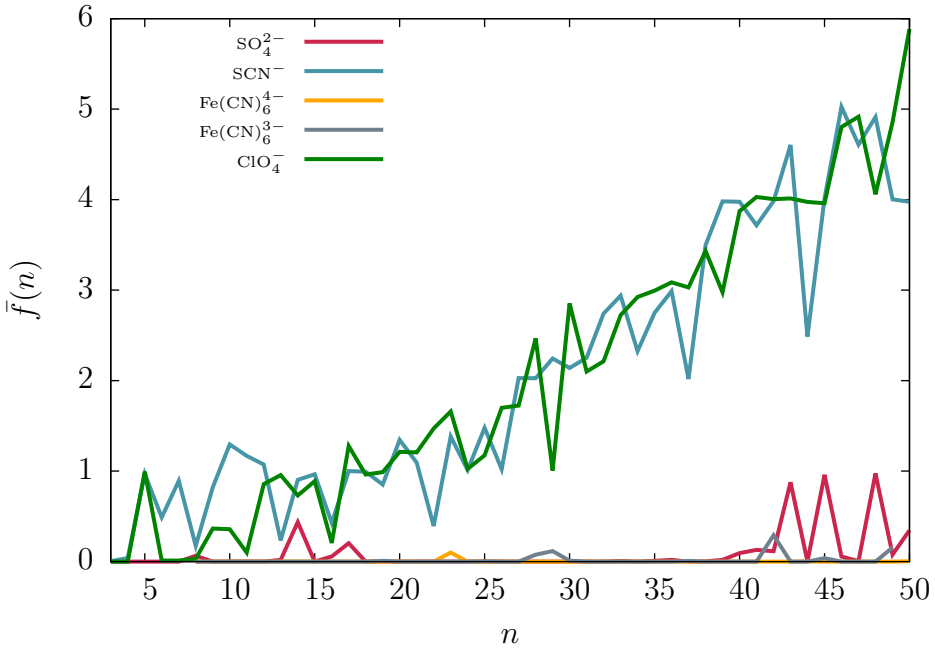


Figure 6.27: Boltzmann-weighted mean number of dangling OH bonds, $\bar{f}$, as a function of the number of water molecules, n, for each of the five hydrated ion systems reported in this thesis, with $3 \leq n \leq 50$.

clusters containing more than 60 water molecules, which is greater than the number of water molecules that it is currently possible to sample sufficiently. A similar result is also obtained from the simulation of the hydrated ferrocyanide cluster, with dangling bonds suppressed for all sizes investigated. In all three cases, the ion is capable of significantly patterning the water structure about itself, resulting in water structures quite distinct from those observed in pure water clusters.

The second regime includes the two singly charged ions, thiocyanate and perchlorate. For both of these ions, dangling OH bonds appear at comparatively small numbers of water molecules (both exhibit dangling OH bonds in their respective $n = 5$ global minimum). In both systems, the number of dangling OH bonds grows steadily and at approximately the same rate. Given that the two ions impart quite different structures to their surrounding water molecules, it is perhaps unusual that they should be so similar. Unfortunately, data for the ensemble of low energy isomers of pure water clusters global minima are not available (although their global minimum structures are reported, they cannot be used to make a direct comparison). It is thus not clear if the thiocyanate and perchlorate ions had reached the pure water limit in regards to number of dangling OH bonds. However, given that even for

small water clusters the number of dangling bonds is typically large ($n = 10$ TIP4P water global minimum has four dangling OH bonds [160]), it is probable that they have not.

In Chapter 5, the concept of the acceptor-donor ratio was introduced in an attempt to relate the hydrated ion micro-structure to the size dependent suppression of dangling OH bonds. It was postulated that by acting as a net *acceptor* of hydrogen bonds, the sulfate ion allows the water molecules to donate more hydrogen bonds than the have to accept, and thus provide more energetically favourable sites with which to hydrogen bond. As the system grows larger, the number of hydrogen bonds to be donated grows as $2n$, but the number of hydrogen bonds the sulfate can accept remains constant (around 12-14 when fully coordinated), and so the effect of the sulfate diminishes as required. The acceptor-donor ratio for cluster i is defined as $r_i = a_i/d_i$, where a_i and d_i are the number of hydrogen bonds accepted and donated by the water molecules in the cluster, respectively.

In Figure 6.28 the ratio of the Boltzmann-weighted mean acceptor-donor ratio, $\bar{r}_{ad}$, is plotted as a function of the number of water molecules for each of the five hydrated ion systems (see Equation 5.3 in Chapter 5). The evolution of $\bar{r}_{ad}$ as a function of the number of water

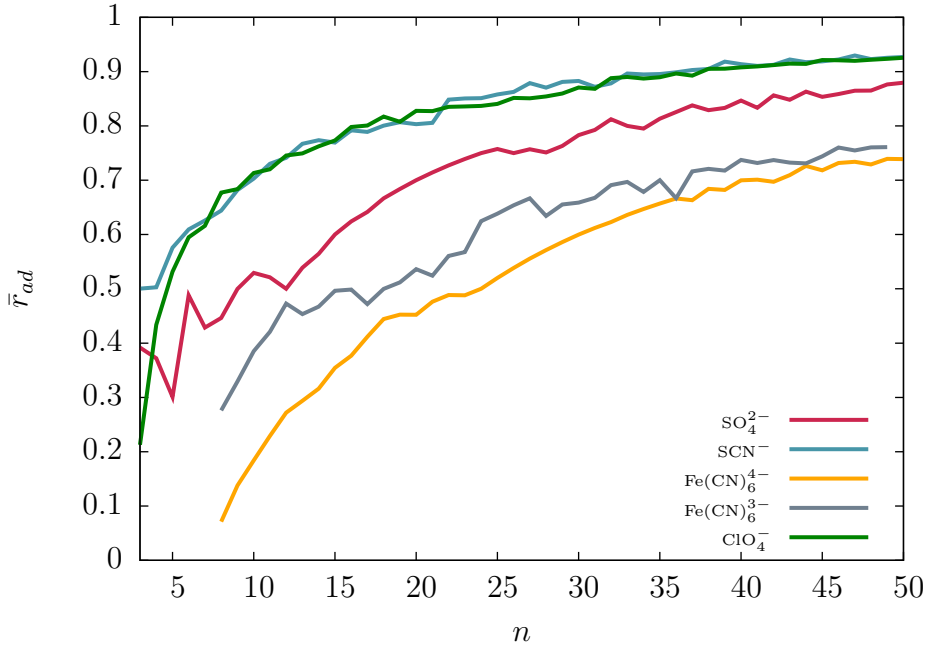


Figure 6.28: Ratio of the Boltzmann-weighted mean number of hydrogen bonds accepted by and donated to a water molecule, $\bar{r}_{ad}$, as a function of the number of water molecules, n, for each of the five hydrated ion systems reported in this thesis, with $3 \leq n \leq 50$.

molecules is determined by the overall charge of the system. The greater the charge on

the ion, the less rapidly the rate of ascent of $\bar{r}_{ad}$ is observed to be. As with the dangling OH bond count, the thiocyanate and perchlorate ions display almost identical behaviour, especially at mid-to-large n. Again, it is not clear why the charge of the ion should play such a dominant role, especially given how unlike the water structure is in the two different systems. Ion geometry plays almost no role in determining the behaviour of $\bar{r}_{ad}$, the sulfate and perchlorate ion are different, as are the ferricyanide and ferrocyanide ions.

6.6 Conclusions

Micro-hydrated ions offer an interesting application of energy landscape theory to study the influence of the ions on local water structure. Rigid-body molecular models interacting via pair-wise empirical potentials are capable of producing a wide range of water structure morphologies, with only slight differences in charge between models dramatically altering the manner in which water molecules organise themselves about the ion. Reoptimising low energy isomers of the thiocyanate ion at the quantum chemical level gives reasonable agreement between the empirical potential and the higher level DFT calculations, though the empirical potential over estimates the strength of the water-thiocyanate interaction, particularly with the sulfur atom. IRPD spectra of size-selected hydrated ferricyanide clusters suggest that dangling OH bonds are suppressed by the ion until approximately 60-70 water molecules, which simulation results are provisionally consistent with. The different searchability of each system is investigated, revealing that the sulfate ion is the easiest to search, whilst the perchlorate is the most challenging. It is suggested that the ease with which a system can be searched is not dependant on the landscape topography, but instead on how modular the surrounding water structure is, and how much compromise is present. This is not necessarily a contradiction; a non-modular or compromised system may require a particularly concerted or collective move to locate an energetically favourable structure, even if the barrier separating it from its neighbouring states is relatively low. Finally, overall charge, not geometry or water structure is shown to be the determining factor in structural measures such as mean number of dangling OH bonds per cluster and the acceptor-donor ratio.

Chapter 7

Conclusions

In this thesis, two quite distinct topics of interest in chemical science have been pursued using a single unifying concept of energy landscape theory. In Part I, two model proteins have been studied using a combination of stationary point sampling and dimensionality reduction methods, and in Part II, the results of an investigation into microsolvation properties of a number of small molecular anions are presented. In this chapter, final general conclusions are drawn and an outlook for future work is suggested and contemplated.

The metric disconnectivity graph analysis in Part I demonstrated that an appropriately chosen order parameter can illuminate relationships between the kinetic structure of a potential surface (i.e. funnels, frustration etc.) with the geometric structure of its constituent parts (i.e. chain slips and twists in the case of the BLN-69 and Gō-69 model proteins). For example, the first principal component of a distribution of low energy minima on the BLN-69 energy landscape is capable of distinguishing between different kinetic funnels. Therefore, this particular coordinate (which corresponds to the two free ends of the protein sliding past one another) is able to demonstrate that in some meaningful way, isomers within a funnel are more structurally similar to each other than to isomers occupying a different funnel. In spite of the apparent success of the metrics proposed, there are still shortcomings which have to be overcome or circumvented. The fraction of native contacts and RMSd metrics relied upon having prior knowledge of the system, and are only applicable to biomolecules such as proteins. The PCA metric is capable of studying systems without resorting to chemical

intuition, but still assumes that the point cloud is approximately linear, and cannot be directly implemented for angular co-ordinates. Isomap allows one to discern low dimensional, non-linear manifolds in the data, and does not make the same assumptions of linearity as PCA. This is clearly a successful strategy, with Isomap distinguishing between the different kinetic structures of the BLN-69 and Gō-69 landscapes. For both of the machine learning metrics implemented, all structures are considered equal, regardless of their energy, which leads to situations such as with Gō-69, where all the variance in a collective co-ordinate is provided by a small number of high energy, unstable minima. Neither the PCA nor Isomap metrics can work for systems containing indistinguishable particles, rendering them unsuitable for a variety of chemical systems. A useful feature of PCA is the ease with which one can project the principal components back into the original space, making it possible to visualise what these directions correspond to. In principle, it should be possible to do the same with Isomap, projecting the approximate geodesics of the manifold back into the original space.

Future work on the metric disconnectivity graph analysis should focus on three interconnected avenues of research: extending the method to more realistic systems, developing different metrics and exploring further uses for collective variables. These topics cannot be explored in isolation, and are not trivial. Tentative work has already begun in this direction, with metric disconnectivity analysis having been performed on databases of minima for a number of head-to-tail cyclic tetrapeptide systems. Cyclic peptides are an interesting class of biomolecules which are capable of aggregating into nano tubes when introduced to lipid environments [202, 203, 204]. A disconnectivity graph of one of the peptides studied, cyclo-[ProβAlaLysβAla], is presented in Figure 7.1. Although there are some initially encouraging results, plenty of technical and conceptual challenges remain. For example, determining a suitable metric has proven to be difficult. A number of metrics for studying cyclic peptides have been trialled, including mean separation of oxygen atoms in the peptide backbone, the aspect ratio of the ellipse formed by the peptide backbone and the Cremer-Pople ring puckering coordinates [205], none of which has proved to be effective. Alternatively, collective co-ordinates could be used as a progress variable in dynamical simulation. By calculating a collective variable from a distribution of minima, one could then project molecular dynamics

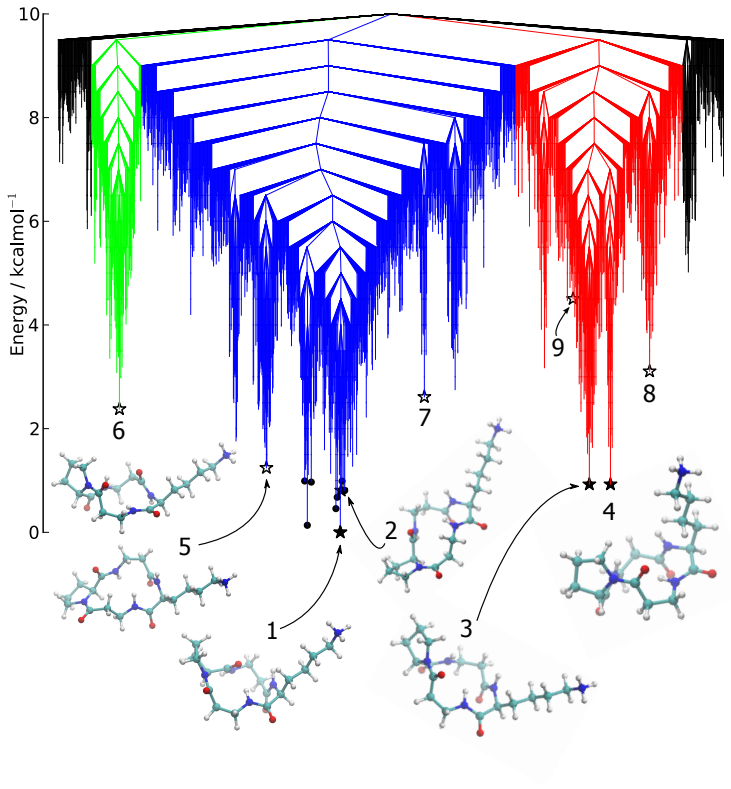


Figure 7.1: Disconnectivity graph of cyclo-[ProβAlaLysβAla], $E_t = 10.0$kcal mol^{-1}, $N_{sp} = 1514$. Colouring used to highlight distinct funnels.

information onto the variable and use it to monitor the simulation. This could prove particularly useful if studying a system which is too large for direct energy landscape exploration, but which contains sub-units which themselves could be individually sampled. A candidate system might be, for example, an aggregation of cyclic peptides monomers self-assembling into a nanotube. In Figure 7.2, the results of a number of independent molecular dynamics simulation of cyclo-[ProβAlaLysβAla] at 300 K in a 16 Å octahedral box containing 2759 TIP3P water molecules starting from different initial structures are projected onto two collective variables computed from a metric disconnectivity graph analysis. Though interesting and worthy of further attention, this work remains very much in its infancy.

In Part II, five different hydrated anion clusters were investigated using a combination of global optimisation and discrete path sampling. Using rigid-body molecular models which interact via pair-wise empirical potentials, a wide range of water structure morphologies can be observed. Only slight differences in charge between models dramatically alter the manner and extent to which water molecules organise themselves about the ion, even if they share the same geometry. For the case of the hydrated sulfate and ferricyanide ions, rigid-body

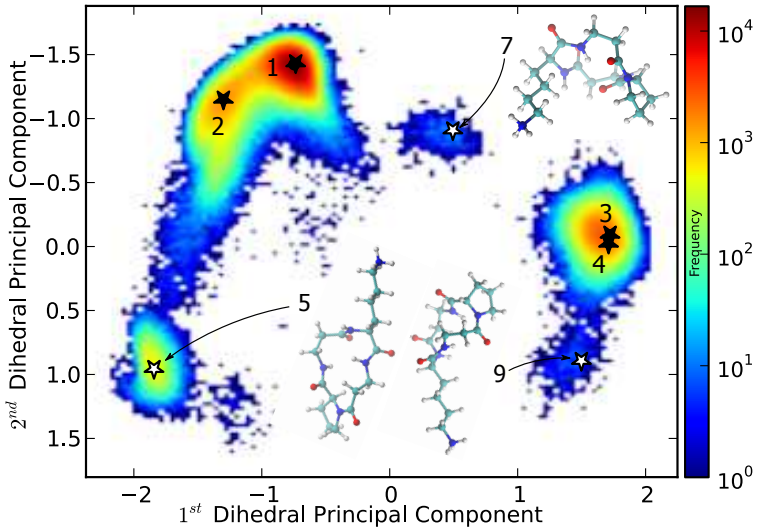


Figure 7.2: Molecular dynamics simulation of a cyclo-[ProβAlaLysβAla] molecule in a 16 Å octahedral box containing 2759 TIP3P water molecules at 300 K. The trajectory is projected onto the first and second dihedral principal components of the peptide backbone.

modelling of the clusters is capable of replicating the physical chemistry of these systems, with both systems exhibiting a size-dependent suppression of dangling OH bonds, provisionally consistent with IRPD spectroscopy of size selected hydrated ions. For the other three ions (thiocyanate, perchlorate and ferrocyanide) no experimental data are currently reported in the literature. However, they are still interesting, as they can be used to probe the effect of differing charge and geometry of the ion on the surrounding water structure. Also, given that the five anions all sit at various points along the Hofmeister series, they can be used to investigate the effect of ion induced solvent ordering in rationalising their Hofmeister position.

A feature common to the smaller ion-water clusters (sulfate, perchlorate, thiocyanate) is the large number of isomers which share an oxygen-ion skeleton but which have different hydrogen bond networks. A global minimum structure will often have a number of energetically competitive hydrogen network isomers which will be difficult using conventional molecular displacements, such as translations and rotations. In order to explore these isomers efficiently, a cycle inversion move class which acts to permute the hydrogen bond network whilst leaving the oxygen-ion skeleton unperturbed was implemented to great success. For the larger ferri-and-ferrocyanide ions, the lack of water-water interactions made direct hydrogen bond network exploration difficult, and the cycle inversion moves were not employed.

The searchability of each system is also presented, revealing that the sulfate ion is the most readily searchable, whilst the perchlorate is the most challenging. The ease with which a system can be searched is not dependant on the landscape topography, but instead on how modular the surrounding water structure are, and how much compromise is present. This is not necessarily a contradiction; a non-modular or compromised system may require a particularly concerted or collective move to locate an energetically favourable structure (thus making it improbable to occur), even if the barrier separating it from its neighbouring states is relatively low.

Cluster geometries minimised at the ab initio level are presented for the sulfate and thiocyanate ions, which were determined by re optimising the lowest energy empirical structures at the DFT level. In both systems the agreement between the empirical and quantum chemical calculations are acceptable, both in terms of the geometries of the minima and their energetic ordering. For the sulfate ion, the rigid-body potential under-estimates the stabilising effect of four-membered rings, whilst for the thiocyanate, the empirical potential over-estimates the strength of the water-thiocyanate interaction, particularly with the sulfur atom.

Future work on this topic should begin to move away from the rigid modelling presented here (as successful as it has been), and begin to concentrate on more chemically faithful modelling methods. This will allow a greater range of problems to be investigated (infrared spectra, UV-vis spectra, quantum tunnelling etc), and produce more accurate geometries and energies. Empirical potentials which permit flexible bonds could be investigated, such as the polarisable AMOEBA potential for water [206], which has already been used to study small hydrated sulfate clusters [140]. It is now increasingly possible to explore the landscapes chemical systems of appreciable size directly at the quantum chemical landscape. For example, pool-Genetic Algorithms are capable of efficiently utilising a large amount of the resources available in modern high performance computing facilities to explore systems containing up to 20 heavy atoms [207, 208].

Appendix A

List of Publications

1. **Visualizing Energy Landscapes with Metric Disconnectivity Graphs**

   L. C. Smeeton, M. T. Oakley, and R. L. Johnston

   J. Comput. Chem., **2014**, 35, 1481-1490

2. **Structures and Energy Landscapes of Hydrated Sulfate Clusters**

   L. C. Smeeton, J. D. Farrell, M. T. Oakley, D. J. Wales, and R. L. Johnston

   J. Chem. Theo. Comp. **2015**, 11, 2377-2384

Bibliography

[1] Wales, D., *Energy Landscapes: Applications to Clusters, Biomolecules and Glasses*, Cambridge Molecular Science, Cambridge University Press, Cambridge, UK, 2003.

[2] Born, M. and Oppenheimer, R., *Annalen der Physik*, 1927, **389**(20), 457–484.

[3] Szabo, A. and Ostlund, N. S., *Modern Quantum Chemistry*, McGraw Hill, New York, USA, 1982.

[4] Kohanoff, J., *Electronic Structure Calculations for Solids and Molecules: Theory and Computational Methods*, Cambridge University Press, Cambridge, UK, 2006.

[5] Alder, B. J. and Wainwright, T. E., *J. Chem. Phys.*, 1959, **31**(2), 459–466.

[6] Rahman, A., *Phys. Rev.*, 1964, **136**(2A), A405–A411.

[7] Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M. N.; Teller, A. H. and Teller, E., *J. Chem. Phys.*, 1953, **21**, 1087–1092.

[8] Murrell, J. N. and Laidler, K. J., *Trans. Faraday Soc.*, 1968, **64**, 371–377.

[9] Mehta, D.; Chen, T.; Hauenstein, J. D. and Wales, D. J., *J. Chem. Phys.*, 2014, **141**(12), 121104.

[10] Wales, D. J. and Berry, R. S., *J. Chem. Soc. Faraday Trans.*, 1992, **88**(4), 543–544.

[11] Karplus, M., *Nat. Chem. Biol.*, 2011, **7**, 401–405.

[12] Oganov, A. R. and Valle, M., *J. Chem. Phys.*, 2009, **130**, 104504.

[13] Sulkowska, J. I.; Noel, J. K. and Onuchic, J. N., *Proc. Natl. Acad. Sci.*, 2012, **109**(44), 17783–17788.

[14] Debenedetti, P. G. and Stillinger, F. H., *Nature*, 2001, **410**, 259–67.

[15] de Souza, V. K. and Wales, D. J., *J. Chem. Phys.*, 2008, **129**, 164507.

[16] Head, J. D.; Weiner, B. and Zerner, M. C., *Int. J. Quantum Chem.*, 1988, **33**(3), 177–186.

[17] Schlegel, H. B., *J. Comput. Chem.*, 2003, **24**(12), 1514–1527.

[18] Schlegel, H. B., *Wiley Interdiscip. Rev.: Comput. Mol. Sci.*, 2011, **1**(5), 790–809.

[19] Wales, D. J., GMIN: A program for finding global minima and calculating thermodynamic properties from basin-sampling. Available at http://www-wales.ch.cam.ac.uk/software.html.

[20] Wales, D. J., OPTIM: A program for characterising stationary points and reaction pathways. Available at http://www-wales.ch.cam.ac.uk/software.html.

[21] Wales, D. J., PATHSAMPLE: A program for refining and analysing kinetic transition networks. Available at http://www-wales.ch.cam.ac.uk/software.html.

[22] Python Energy Landscape Explorer. Available at http://www.github.com/pele-python/pele.

[23] Valiev, M.; Bylaska, E. J.; Govind, N.; Kowalski, K.; Straatsma, T. P.; Van Dam, H. J. J.; Wang, D.; Nieplocha, J.; Apra, E.; Windus, T. L. and de Jong, W. A., *Comput. Phys. Commun.*, 2010, **181**(9), 1477–1489.

[24] Broyden, C. G., *J. Inst. Math. Its Appl.*, 1970, **6**(1), 76–90.

[25] Fletcher, R., *Comput. J.*, 1970, **13**(3), 317–322.

[26] Goldfarb, D., *Math. Comput.*, 1970, **24**(109), 23–26.

[27] Shanno, D. F., *Math. Comput.*, 1970, **24**(111), 647–656.

[28] Liu, D. C. and Nocedal, J., *Math. Program.*, 1989, **45**(3), 503–528.

[29] More, J. J.; Thuente, D. J. and Mcs-p, P., *ACM Trans. Math. Software*, 1992, **20**, 286–307.

[30] Trygubenko, S. A. and Wales, D. J., *J. Chem. Phys.*, 2004, **120**(5), 2082–2094.

[31] Munro, L. J. and Wales, D. J., *Phys. Rev. B*, 1999, **59**, 3969–3980.

[32] Henkelman, G.; Jónsson, H. and Jonsson, H., *J. Chem. Phys.*, 1999, **111**(15), 7010–7022.

[33] Elber, R. and Karplus, M., *Chem. Phys. Lett.*, 1987, **139**(5), 375–380.

[34] Czerminski, R. and Elber, R., *J. Chem. Phys.*, 1990, **92**, 5580–5601.

[35] Jonsson, H.; Mills, G. and Jacobsen, K. W., *Classical and Quantum Dynamics in Condensed Phase Simulations - Proceedings of the International School of Physics*, 1998, pages 385–404.

[36] Henkelman, G.; Uberuaga, B. P. and Jónsson, H., *J. Chem. Phys.*, 2000, **113**(22), 9901–9904.

[37] Sheppard, D.; Terrell, R. and Henkelman, G., *J. Chem. Phys.*, 2008, **128**(2008).

[38] Henkelman, G. and Jonsson, H., *J. Chem. Phys.*, 2000, **113**(22), 9978.

[39] Kumeda, Y.; Wales, D. J. and Munro, L. J., *Chem. Phys. Lett.*, 2001, **34**, 185–194.

[40] Pancí, J., *Collect. Czechoslov. Chem. Commun.*, 1975, **40**(4), 1112–1118.

[41] Cerjan, C. J., *J. Chem. Phys.*, 1981, **75**(6), 2800.

[42] Wales, D. J., *J. Chem. Phys.*, 1994, **101**(5), 3750.

[43] Wales, D. J. and Walsh, T. R., *J. Chem. Phys.*, 1996, **105**(16), 6957.

[44] Stillinger, F. H. and We, T. A., *Science*, 1984, **225**(4666), 983–989.

[45] Stillinger, F. H. and Weber, T. A., *Phys. Rev. A*, 1982, **25**(2), 978–989.

[46] Stillinger, F., *Phys. Rev. E*, 1999, **59**(1), 48–51.

[47] Li, Z. and Scheraga, H. A., *Proc. Natl. Acad. Sci. U.S.A*, 1987, **84**(19), 6611–6615.

[48] Wales, D. J. and Doye, J. P. K., *J. Phys. Chem. A*, 1997, **101**(28), 5111–5116.

[49] Wales, D. J. and Scheraga, H. A., *Science*, 1999, **285**(5432), 1368–1372.

[50] Becker, O. M., *Proteins*, 1997, **27**, 213–226.

[51] García, A. and Blumenfeld, R., *Phys. D*, 1997, **107**, 225–239.

[52] Troyer, J. M. and Cohen, F. E., *Proteins*, 1995, **23**, 97–110.

[53] Becker, O. M. and Karplus, M., *J. Chem. Phys.*, 1997, **106**(4), 1495.

[54] Hoffmann, K. H. and Sibani, P., *Phys. Rev. A*, 1988, **38**, 4261–4270.

[55] Wales, D. J.; Miller, M. A. and Walsh, T. R., *Nature*, 1998, **394**, 758–760.

[56] Middleton, T.; Hernández-Rojas, J.; Mortenson, P. and Wales, D., *Phys. Rev. B*, 2001, **64**, 184201.

[57] Wales, D. J., *Curr. Opin. in Struct. Biol.*, 2010, **20**, 3–10.

[58] Wales, D. J. and Bogdan, T. V., *J. Phys. Chem. B*, 2006, **110**, 20765–20776.

[59] Krivov, S. V. and Karplus, M., *J. Chem. Phys.*, 2002, **117**, 10894.

[60] Evans, D. A. and Wales, D. J., *J. Chem. Phys.*, 2003, **118**, 3891–3897.

[61] Evans, D. A. and Wales, D. J., *J. Chem. Phys.*, 2003, **119**, 9947–9955.

[62] Oakley, M. T.; Oheix, E.; Peacock, A. F. A. and Johnston, R. L., *J. Phys. Chem. B*, 2013, **117**, 8122–8134.

[63] Lempesis, N.; Boulougouris, G. C. and Theodorou, D. N., *J. Chem. Phys.*, 2013, **138**, 12A545.

[64] Sibani, P. and Schön, J. C., *Euro. Phys. Lett.*, 1993, **22**, 479–485.

[65] Komatsuzaki, T.; Hoshino, K.; Matsunaga, Y.; Rylance, G. J.; Johnston, R. L. and Wales, D. J., *J. Chem. Phys.*, 2005, **122**, 84714–87423.

[66] Rylance, G. J.; Johnston, R. L.; Matsunaga, Y.; Li, C.-B.; Baba, A. and Komatsuzaki, T., *Proc. Natl. Acad. Sci.*, 2006, **103**, 18551–18555.

[67] Hunter, J. D., *Comput. Sci. Eng.*, 2007, **9**, 90–95.

[68] Pérez, F. and Granger, B. E., *Comput. Sci. Eng.*, 2007, **9**, 21–29.

[69] Miller, M. A., DISCONNECT: A program for producing disconnectivity graphs, 2013.

[70] Frauenfelder, H.; Chan, S. S. and Chan, W. S., *The Physics of Proteins*, Springer, New York, USA, 2010.

[71] Pauling, L.; Corey, R. B. and Branson, H. R., *Proc. Natl. Acad. Sci.*, 1951, **37**, 205–211.

[72] Schulz, G. E. and Schirmer, R. H., *Principles of Protein Structure*, Springer, New York, USA, 1979.

[73] Huang, K., *Lectures on Statistical Physics and Protein Folding*, World Scientific, Singapore, 2005.

[74] Dill, K. A., *Biochemistry*, 1990, **29**, 7133–7155.

[75] Anfinsen, C. B.; Haber, E.; Sela, M. and Jr., F. H. W., *Proc. Natl. Acad. Sci.*, 1961, **47**, 1309–1314.

[76] Hartl, F.-U.; Hlodan, R. and Langer, T., *Trends Biochem. Sci.*, 1994, **19**, 20–25.

[77] Hartl, F. U., *Philos. Trans. Biol. Sci.*, 1995, **348**, 107–112.

[78] Levinthal, C. In *Mossbauer Spectrosc. Biol. Syst. Proc. a Meet. held Allert. House, Monticello, Illinois.*, pages 22–24. University of Illinois Press, 1969.

[79] Zwanzig, R.; Szabo, A. and Bagchi, B., *Proc. Natl. Acad. Sci.*, 1992, **89**, 20–22.

[80] Mirny, L. and Shakhnovich, E., *Annu. Rev. Biophys. Biomol. Struct.*, 2001, **30**, 369–396.

[81] Marrink, S. J.; Risselada, H. J.; Yefimov, S.; Tieleman, D. P. and de Vries, A. H., *J. Phys. Chem. B*, 2007, **111**, 7812–7824.

[82] Honeycutt, J. D. and Thirumalai, D., *Proc. Natl. Acad. Sci.*, 1990, **87**, 3526–3529.

[83] Thirumalai, D. and Guo, Z., *Biopolymers*, 1995, **35**, 137–140.

[84] Berry, R. S.; Elmaci, N.; Rose, J. P. and Vekhter, B., *Proc. Natl. Acad. Sci.*, 1997, **94**, 9520–9524.

[85] Bryngelson, J. D.; Onuchic, J. N.; Socci, N. D. and Wolynes, P. G., *Proteins*, 1995, **21**, 167–195.

[86] Guo, Z. and Brooks, C. L., *Biopolymers*, 1997, **42**(7), 745–57.

[87] Onuchic, J. N.; Luthey-Schulten, Z. and Wolynes, P. G., *Annu. Rev. Phys. Chem.*, 1997, **48**, 545–600.

[88] Kim, S.-Y., *J. Chem. Phys.*, 2010, **133**, 135102.

[89] Oakley, M. T.; Wales, D. J. and Johnston, R. L., *J. Phys. Chem. B*, 2011, **115**, 11525–11529.

[90] Ueda, Y.; Taketomi, H. and Gō, N., *Biopolymers*, 1978, **17**, 1531–1548.

[91] Kim, J. and Keyes, T., *J. Phys. Chem. B*, 2008, **112**, 954–966.

[92] Lechner, W. and Dellago, C., *J. Chem. Phys.*, 2008, **129**(11), 114707.

[93] Wang, J.; Oliveira, R. J.; Chu, X.; Whitford, P. C.; Chahine, J.; Han, W.; Wang, E.; Onuchic, J. N. and Leite, V. B. P., *Proc. Natl. Acad. Sci.*, 2012, **109**, 15763–16768.

[94] Kabsch, W., *Acta Crystallogr., Sect. A*, 1978, **34**, 827–828.

[95] Shlens, J. In *Systems Neurobiology Laboratory, Salk Institute for Biological Studies*, 2005.

[96] Riccardi, L.; Nguyen, P. H. and Stock, G., *J. Chem. Theory Comput.*, 2012, **8**, 1471–1479.

[97] Altis, A.; Nguyen, P. H.; Hegger, R. and Stock, G., *J. Chem. Phys.*, 2007, **126**, 244111.

[98] McLachlan, A., *Biopolymers*, 1984, **23**, 1325–1331.

[99] Tenenbaum, J. B.; de Silva, V. and Langford, J. C., *Science*, 2000, **290**, 2319–2323.

[100] Das, P.; Moll, M.; Stamati, H.; Kavraki, L. E. and Clementi, C., *Proc. Natl. Acad. Sci.*, 2006, **103**, 9885–9890.

[101] Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M. and Duchesnay, E., *J. Mach. Learn. Res.*, 2011, **12**, 2825–2830.

[102] Smeeton, L. C.; Oakley, M. T. and Johnston, R. L., *J. Comput. Chem.*, 2014, **35**(20), 1481–1490.

[103] Ceriotti, M.; Tribello, G. A. and Parrinello, M., *J. Chem. Theory Comput.*, 2013, **9**, 1521–1532.

[104] Rohrdanz, M. A.; Zheng, W.; Maggioni, M. and Clementi, C., *J. Chem. Phys.*, 2011, **134**, 124116.

[105] Cazals, F.; Chazal, F. and Giesen, J. In *Nonlinear Computational Geometry*, Emiris, I. Z.; Sottile, F. and Theobald, T., Eds., Vol. 151 of *The IMA Volumes in Mathematics and its Applications;* Springer New York, 2010; pages 1–34.

[106] Dorogovstev, S. N., *Lectures on Complex Networks*, Oxford University Press, 2010.

[107] Cazals, F.; Dreyfus, T.; Mazauric, D.; Roth, C.-A. and Robert, C. H., *J. Comput. Chem.*, 2015, **36**(16), 1213–1231.

[108] Kebarle, P., *J. Chem. Phys.*, 1968, **49**(2), 817.

[109] Castleman, A. W. and Keesee, R. G., *Chem. Rev.*, 1986, **86**(3), 589–618.

[110] Robertson, W. H. and Johnson, M. A., *Annu. Rev. Phys. Chem.*, 2003, **54**, 173–213.

[111] Yamashita, M. and Fenn, J. B., *J. Phys. Chem.*, 1984, **88**(20), 4451–4459.

[112] Magnera, T. F.; David, D. E.; Stulik, D.; Orth, R. G.; Jonkman, H. T. and Michl, J., *J. Am. Chem. Soc.*, 1989, **111**(14), 5036–5043.

[113] Achatz, U.; Joos, S.; Berg, C.; Beyer, M.; Niedner-Schatteburg, G. and Bondybey, V., *Chem. Phys. Lett.*, 1998, **291**(5-6), 459–464.

[114] Cabarcos, O. M.; Weinheimer, C. J.; Lisy, J. M. and Xantheas, S. S., *J. Chem. Phys.*, 1999, **110**(1), 5.

[115] Ayotte, P.; Weddle, G. H.; Kim, J. and Johnson, M. A., *Chem. Phys.*, 1998, **239**(1-3), 485–491.

[116] Ayotte, P.; Nielsen, S. B.; Weddle, G. H.; Johnson, M. A. and Xantheas, S. S., *J. Phys. Chem. A*, 1999, **103**(50), 10665–10669.

[117] Combariza, J. E.; Kestner, N. R. and Jortner, J., *J. Chem. Phys.*, 1994, **100**(4), 2851.

[118] Xantheas, S. S., *J. Phys. Chem.*, 1996, **100**(23), 9703–9713.

[119] Perera, L. and Berkowitz, M. L., *J. Chem. Phys.*, 1993, **99**(5), 4222.

[120] Perera, L. and Berkowitz, M. L., *J. Chem. Phys.*, 1994, **100**(4), 3085.

[121] Koneshan, S.; Rasaiah, J. C.; Lynden-Bell, R. M. and Lee, S. H., *J. Phys. Chem. B*, 1998, **102**(21), 4193–4204.

[122] Lide, D. R., *CRC Handbook of Chemistry and Physics, 94th Edition, 2013-2014*, Vol. 53, 2013.

[123] Boxford, W. E. and Dessent, C. E. H., *Phys. Chem. Chem. Phys.*, 2006, **8**(44), 5151–5165.

[124] Wang, X.-B.; Ding, C.-F. and Wang, L.-S., *Phys. Rev. Lett.*, 1998, **81**(16), 3351–3354.

[125] Wang, X.-B. and Wang, L.-S., 1999, **400**(6741), 245–248.

[126] Wang, X.-B. B.; Nicholas, J. B. and Wang, L.-S. S., *J. Chem. Phys.*, 2000, **113**(24), 10837–10840.

[127] Wang, X.-B.; Ding, C.-F.; Nicholas, J. B.; Dixon, D. A. and Wang, L.-S., *J. Phys. Chem. A*, 1999, **103**(18), 3423–3429.

[128] Friedrich, J.; Gilb, S.; Ehrler, O. T.; Behrendt, A. and Kappes, M. M., *J. Chem. Phys.*, 2002, **117**(6), 2635.

[129] Wang, X. B.; Yang, X.; Nicholas, J. B. and Wang, L. S., *Science*, 2001, **294**(5545), 1322–1325.

[130] DiTucci, M. J.; Heiles, S. and Williams, E. R., *J. Am. Chem. Soc.*, 2015, **137**(4), 1650–1657.

[131] Garand, E.; Wende, T.; Goebbert, D. J.; Bergmann, R.; Meijer, G.; Neumark, D. M. and Asmis, K. R., *J. Am. Chem. Soc.*, 2010, **132**(2), 849–856.

[132] Yacovitch, T. I.; Wende, T.; Jiang, L.; Heine, N.; Meijer, G.; Neumark, D. M. and Asmis, K. R., *J. Phys. Chem. Lett.*, 2011, **2**(17), 2135–2140.

[133] Lee, H. M.; Kim, D.; Singh, N. J.; Koaski, M. and Kim, K. S., *J. Chem. Phys.*, 2007, **127**(16), 164311.

[134] Lee, H. M.; Tarkeshwar, P. and Kim, K. S., *J. Chem. Phys.*, 2004, **121**(10), 4657–64.

[135] Ramanathan, V.; Crutzen, P. J.; Kiehl, J. T. and Rosenfeld, D., *Science*, 2001, **294**(5549), 2119–2124.

[136] Bush, M. F.; Saykally, R. J. and Williams, E. R., *J. Am. Chem. Soc.*, 2007, **129**(8), 2220–2221.

[137] Yang, X.; Wang, X.-B. and Wang, L.-S., *J. Phys. Chem. A*, 2002, **106**(33), 7607–7616.

[138] Wan, Q.; Spanu, L. and Galli, G., *J. Phys. Chem. B*, 2012, **116**(31), 9460–9466.

[139] Jungwirth, P.; Curtis, J. E. and Tobias, D. J., *Chem. Phys. Lett.*, 2003, **367**(5-6), 704–710.

[140] Lambrecht, D. S.; Clark, G. N. I.; Head-Gordon, T. and Head-Gordon, M., *J. Phys. Chem. A*, 2011, **115**(41), 11438–11454.

[141] O'Brien, J. T.; Prell, J. S.; Bush, M. F. and Williams, E. R., *J. Am. Chem. Soc.*, 2010, **132**(24), 8248–8249.

[142] Smeeton, L. C.; Farrell, J. D.; Oakley, M. T.; Wales, D. J. and Johnston, R. L., *J. Chem. Theo. Comp.*, 2015, **11**(5), 2377–2384.

[143] Kunz, W.; Lo Nostro, P. and Ninham, B., *Curr. Opin. Colloid Interface Sci.*, 2004, **9**(1-2), 1–18.

[144] Hofmeister, F., *Archiv für Experimentelle Pathologie und Pharmakologie*, 1888, **24**(4-5), 247–260.

[145] Pegram, L. M. and Record, M. T., *J. Phys. Chem. B*, 2007, **111**(19), 5411–5417.

[146] Baldwin, R. L., *Biophys. J.*, 1996, **71**(4), 2056–2063.

[147] Russo, D., *Chem. Phys.*, 2008, **345**(2-3), 200–211.

[148] Levy-Sakin, M.; Scherzer-Attali, R. and Gazit, E., In *Protein and Peptide Folding, Misfolding, and Non-Folding*, John Wiley & Sons, Inc., 2012; pages 441–478.

[149] Zhang, Y. and Cremer, P. S., *Curr. Opin. Chem. Biol.*, 2006, **10**(6), 658–663.

[150] Smith, J. D.; Saykally, R. J. and Geissler, P. L., *J. Am. Chem. Soc.*, 2007, **129**(45), 13847–13856.

[151] Freire, M. G.; Neves, C. M. S. S.; Silva, A. M. S.; Santos, L. M. N. B. F.; Marrucho, I. M.; Rebelo, L. P. N.; Shah, J. K.; Maginn, E. J. and Coutinho, J. A. P., *J. Phys. Chem. B*, 2010, **114**(5), 2004–2014.

[152] Uejio, J. S.; Schwartz, C. P.; Duffin, A. M.; Drisdell, W. S.; Cohen, R. C. and Saykally, R. J., *Proc. Natl. Acad. Sci.*, 2008, **105**(19), 6809–6812.

[153] Cacace, M. G.; Landau, E. M. and Ramsden, J. J., *Q. Rev. Biophys.*, 1997, **30**(3), 241–277.

[154] Omta, A. W.; Kropman, M. F.; Woutersen, S. and Bakker, H. J., *Science*, 2003, **301**(5631), 347–349.

[155] Hepler, L. G., *Can. J. Chem.*, 1969, **47**(24), 4613–4617.

[156] Gurau, M. C.; Lim, S.-M.; Castellana, E. T.; Albertorio, F.; Kataoka, S. and Cremer, P. S., *J. Am. Chem. Soc.*, 2004, **126**(34), 10522–10523.

[157] Gopalakrishnan, S.; Jungwirth, P.; Tobias, D. J. and Allen, H. C., *J. Phys. Chem.. B*, 2005, **109**(18), 8861–8872.

[158] Liu, D.; Ma, G.; Levering, L. M. and Allen, H. C., *J. Phys. Chem. B*, 2004, **108**(7), 2252–2260.

[159] O'Brien, J. T. and Williams, E. R., *J. Am. Chem. Soc.*, 2012, **134**(24), 10228–10236.

[160] Wales, D. J. and Hodges, M. P., *Chem. Phys. Lett.*, 1998, **286**(1-2), 65–72.

[161] Chakrabarti, D. and Wales, D. J., *Phys. Chem. Chem. Phys.*, 2009, **11**(12), 1970–1976.

[162] Ruehle, V.; Kusumaatmaja, H.; Chakrabarti, D. and Wales, D. J., *J. Chem. Theory Comput.*, 2013, **9**(9), 4026–4034.

[163] Goldstein, H., *Classical mechanics*, World student series, Addison-Wesley, Amsterdam, 1980.

[164] Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W. and Klein, M. L., *J. Chem. Phys.*, 1983, **79**(2), 926.

[165] Kazachenko, S. and Thakkar, A. J., *Chem. Phys. Lett.*, 2009, **476**(1-3), 120–124.

[166] Takeuchi, H., *J. Chem. Inf. Model.*, 2008, **48**(11), 2226–2233.

[167] Kabrede, H., *Chem. Phys. Lett.*, 2006, **430**(4-6), 336–339.

[168] Hartke, B., *Zeitschrift für Physikalische Chemie*, 2000, **214**(9/2000), 1252.

[169] Kazachenko, S. and Thakkar, A. J., *Mol. Phys.*, 2010, **108**(17), 2187–2193.

[170] Kazimirski, J. K. and Buch, V., *J. Phys. Chem. A*, 2003, **107**(46), 9762–9775.

[171] Kabrede, H. and Hentschke, R., *J. Phys. Chem. B*, 2003, **107**(16), 3914–3920.

[172] Tsai, C. J. and Jordan, K. D., *J. Phys. Chem.*, 1993, **97**(20), 5208–5210.

[173] Kazachenko, S. and Thakkar, A. J., *J. Chem. Phys.*, 2013, **138**(19), 194302.

[174] Sanz, E.; Vega, C.; Abascal, J. L. F. and MacDowell, L. G., *J. Chem. Phys.*, 2004, **121**(2), 1165.

[175] Sanz, E.; Vega, C.; Abascal, J. and MacDowell, L., *Phys. Rev. Lett.*, 2004, **92**(25), 255701.

[176] Abascal, J. L. F. and Vega, C., *J. Chem. Phys.*, 2005, **123**(23), 234505.

[177] Cannon, W. R.; Pettitt, B. M. and McCammon, J. A., *J. Phys. Chem.*, 1994, pages 6225–6230.

[178] Pye, C. C. and Rudolph, W. W., *J. Phys. Chem. A*, 2001, **105**(5), 905–912.

[179] Heinje, G.; Luck, W. A. P. and Heinzinger, K., *J. Phys. Chem.*, 1987, **91**(2), 331–338.

[180] Pye, C. C. and Walker, V. E. J., *J. Phys. Chem.. A*, 2011, **115**(45), 13007–13015.

[181] Chaumont, A. and Wipff, G., *Inorg. Chem.*, 2009, **48**(10), 4277–4289.

[182] Jorgensen, W. L.; Maxwell, D. S. and Tirado-Rives, J., *J. Am. Chem. Soc.*, 1996, **118**(45), 11225–11236.

[183] Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W. and Kollman, P. A., *J. Am. Chem. Soc.*, 1995, **117**(19), 5179–5197.

[184] Prampolini, G.; Yu, P.; Pizzanelli, S.; Cacelli, I.; Yang, F.; Zhao, J. and Wang, J., *J. Phys. Chem.. B*, 2014, **118**(51), 14899–14912.

[185] Cacelli, I. and Prampolini, G., *J. Chem. Theo. Comp.*, 2007, **3**(5), 1803–1817.

[186] Barone, V.; Cacelli, I.; De Mitri, N.; Licari, D.; Monti, S. and Prampolini, G., *Phys. Chem. Chem. Phys.*, 2013, **15**(11), 3736–3751.

[187] Aguilar, C. M.; De Almeida, W. B. and Rocha, W. R., *Chem. Phys. Lett.*, 2007, **449**(1-3), 144–148.

[188] Rappe, A. K.; Casewit, C. J.; Colwell, K. S.; Goddard, W. A. and Skiff, W. M., *J. Am. Chem. Soc.*, 1992, **114**(25), 10024–10035.

[189] Price, M. L. P.; Ostrovsky, D. and Jorgensen, W. L., *J. Comput. Chem.*, 2001, **22**(13), 1340–1352.

[190] Newman, M., *Networks: An Introduction*, Oxford University Press, Inc., New York, NY, USA, 2010.

[191] Hagberg, A. A.; Schult, D. A. and Swart, P. J. In *Proceedings of the 7th Python in Science Conference (SciPy2008)*, pages 11–15, Pasadena, CA USA, 2008.

[192] Mardirossian, N.; Lambrecht, D. S.; McCaslin, L.; Xantheas, S. S. and Head-Gordon, M., *J. Chem. Theory Comput.*, 2013, **9**(3), 1368–1380.

[193] Du, Q.; Superfine, R.; Freysz, E. and Shen, Y., *Phys. Rev. Lett.*, 1993, **70**(15), 2313–2316.

[194] Miller, Y.; Chaban, G. M.; Zhou, J.; Asmis, K. R.; Neumark, D. M. and Gerber, R. B., *J. Chem. Phys.*, 2007, **127**(9), 094305.

[195] Gao, B. and Liu, Z.-f., *J. Chem. Phys.*, 2005, **123**(22), 224302.

[196] Wales, D. J. and Ohmine, I., *J. Chem. Phys.*, 1993, **98**, 7257–7268.

[197] Wales, D. J. and Ohmine, I., *J. Chem. Phys.*, 1993, **98**, 7245–7256.

[198] Heiles, S.; Logsdail, A. J.; Schäfer, R. and Johnston, R. L., *Nanoscale*, 2012, **4**(4), 1109–1115.

[199] Heiles, S. and Johnston, R. L., *Int. J. Quantum Chem.*, 2013, **113**(18), 2091–2109.

[200] Do, H. and Besley, N. A., *J. Chem. Phys.*, 2012, **137**(13), 134106.

[201] Botti, A.; Pagnotta, S. E.; Bruni, F. and Ricci, M. A., *J. Phys. Chem.. B*, 2009, **113**(29), 10014–10021.

[202] Engels, M.; Bashford, D. and Ghadiri, M. R., *Soc, J Am Chem*, 1995, **117**, 9151–9158.

[203] Tarek, M.; Maigret, B. and Chipot, C., *Biophys. J.*, 2003, **85**(4), 2287–2298.

[204] Zhu, J.; Cheng, J.; Liao, Z.; Lai, Z. and Liu, B., *J. Comput. Aided. Mol. Des.*, 2008, **22**, 773–781.

[205] Cremer, D. and Pople, J. A., *J. Am. Chem. Soc.*, 1975, **97**(6), 1354–1358.

[206] Ren, P. and Ponder, J. W., *J. Phys. Chem. B*, 2003, **107**(24), 5933–5947.

[207] Shayeghi, A.; Götz, D.; Davis, J. B. A.; Schäfer, R. and Johnston, R. L., *Phys. Chem. Chem. Phys.*, 2015, **17**(3), 2104–2112.

[208] Davis, J.; Shayeghi, A.; Horswell, S. L. and Johnston, R. L., *Nanoscale*, 2015.